

PhD Thesis

THE ECONOMICS OF
SAVINGS AND LOAN ASSOCIATIONS

EVIDENCE FROM A RANDOMIZED
CONTROL TRIAL IN MALAWI

Ole Dahl Rasmussen

2013

Department of Business and Economics
University of Southern Denmark

THE ECONOMICS OF SAVINGS AND
LOAN ASSOCIATIONS
EVIDENCE FROM A RANDOMIZED
CONTROL TRIAL IN MALAWI

Ole Dahl Rasmussen

2013

A dissertation submitted to the Faculty of Social Sciences
in partial fulfilment of the requirements for the PhD
degree in economics

Supervisor
Nikolaj Malchow-Møller

Co-supervisor
Thomas Barnebeck-Andersen

Department of Business and Economics
University of Southern Denmark

CONTENTS

Preface.....	xi
Indledning og resumé på dansk (Introduction and Summary in Danish).....	15
Introduction and Summary of Chapters.....	27
Chapter 1	41
<i>The Impact of Village Savings and Loan Associations on Household Welfare: Evidence from a Cluster Randomized Trial (with Christopher Ksoll, Helene Bie Lilleør and Jonas Helth Lønborg)</i>	
Chapter 2	143
<i>Can Microfinance Reach the Poorest: Evidence from a Community-Managed Microfinance Intervention (with Jonas Helth Lønborg)</i>	
Chapter 3	187
<i>Small Groups, Large Profits: Calculating Interest Rates in Community-Managed Microfinance</i>	
Chapter 4	221
<i>Walking the Talk: the Need for a Trial Registry for Development Interventions (with Thomas Barnebeck Andersen and Nikolaj Malchow-Møller)</i>	

DETAILED CONTENTS

Preface.....	xi
Indledning og resumé på dansk (Introduction and Summary in Danish).....	15
Lokale strategier i bistandsarbejdet.....	16
Spare-låne-grupper som mikrofinans.....	18
Lodtrækningsforsøg i bstanden.....	20
Resumé af afhandlingen	21
Referencer	23
Introduction and Summary of Chapters.....	27
In Defense of a Local Strategy.....	27
SLAs as Microfinance	29
Making Sense of Randomization in Aid	31
Summary of Chapters.....	33
References	38
Chapter 1	41
<i>The Impact of Village Savings and Loan Associations on Household Welfare: Evidence from a Cluster Randomized Trial (with Christopher Ksoll, Helene Bie Lilleør and Jonas Helth Lønborg)</i>	
1. Introduction	44
2. Intervention and Research Design.....	47
3. Data and Estimation Strategy	67
4. Results on Predefined Outcomes.....	91
5. Intermediate Outcomes.....	102
6. Cost Effectiveness	114
7. Conclusion.....	116
8. Appendix A: Estimating Food Consumption	119
9. References	135

Chapter 2143

Can Microfinance Reach the Poorest: Evidence from a Community-Managed Microfinance Intervention (with Jonas Helth Lønborg)

Introduction	146
The Intervention	148
Targeting and Outreach: A Literature Review.....	149
The Outreach Ratio	153
A Leaking Pipeline	159
The Data	161
Empirical Strategy	165
Results	168
Conclusion.....	179
References	182

Chapter 3187

Small Groups, Large Profits: Calculating Interest Rates in Community-Managed Microfinance

Introduction	190
Are savings groups different?	191
To compound or not to compound	192
Non-standard interest rate calculations	194
Data on interest rates and savings profiles.....	198
How should we report interest rates on savings in savings groups?..	202
Using comparable interest rates for monitoring.....	206
Conclusions	209
Annex 1. The standard financial calculation.....	211
Annex 2. Determining the annual interest rate using the simple return on savings and the age of the group	214
Annex 3. Determining the monthly interest rate using the simple return on savings and the age of the group	217
References	219

Chapter 4221

*Walking the Talk: the Need for a Trial Registry for
Development Interventions (with Thomas Barnebeck
Andersen and Nikolaj Malchow-Møller)*

Introduction	224
The immediate benefits of a trial registry	226
Is a trial registry really needed?	229
Letting researchers research: other benefits of a trial registry	230
Copying the wheel: trial registries in medicine	233
Creating a credibility infrastructure: a trial registry is not enough....	240
What are the important features of a trial registry?	241
Conclusions	243
Appendix 1. Proposal for contents of a trial registry for development interventions	245
References	249

PREFACE

In hindsight, most achievements seem planned and purposeful. Looking back, this PhD thesis is a logical continuation of the work I did in my master's thesis, the inspiration I got from studying development economics with Michael Carter at the University of Wisconsin-Madison in 2006, as well as the years working with microfinance at DanChurchAid, a Danish civil society organization.

But there is more to it. Without at least two coincidences this thesis would almost certainly never have materialized. The first coincidence happened on January 20th, 2009, when I received an e-mail from Jonas Helth Lønborg, whom I had never met. He wrote me to ask if he could do his own PhD degree in collaboration with DanChurchAid. His e-mail led me to discuss the impact evaluation of a newly initiated project in depth with Helene Bie Lilleør at the Rockwool Foundation Research Unit. Not long after, Helene, Jonas and I had ambitious plans for the evaluation that lead to the first chapter of this thesis. But I did not yet have a PhD in mind. This was changed by the second coincidence which took place some three months later, as I was discussing practical microfinance with the then international director of DanChurchAid, Christian Friis Bach. At some point he said: Shouldn't you also pursue a PhD? Without his encouragement and the backing of DanChurchAid, this thesis would still be little more than an idea.

Along the way, I got support from a large number of individuals, whom I would like to thank. My advisors Nikolaj Malchow-Møller and Thomas Barnebeck Andersen have always been available to provide insightful comments on the details of econometric specifications, inputs on the (lack of) logic in my arguments or, perhaps Thomas' favourite, a must-read book on a relevant topic.

In Malawi, a special word of thanks goes to the 1,775 Malawians who have volunteered their valuable time answering our questions. Needless to say, there would be no data and thus no thesis without them or the competent and dedicated staff at the Invest in Knowledge Initiative who conducted the interviews. Waluza Munthali at the local

civil society organisation SOLDEV headed the work of establishing the one hundred savings and loan associations scrutinized on the following pages, a task which was made more cumbersome by our analysis. His professionalism and friendliness is extraordinary and I owe him great thanks. At the DanChurchAid Office in Malawi, I had the pleasure of working with several colleagues, particularly Bernard Kamanga, Lennart Skov-Hansen and Cecilie Winter, who all provided crucial support along the way.

As a part of the Danish Industrial PhD programme, I have been based at DanChurchAid for a large share of my time. Here, a special word of thanks to General Secretary Henrik Stubkjær for DCA's engagement in research and support of this PhD. Also at DCA, I have really enjoyed the friendly, practical and thoughtful atmosphere among my colleagues at the Program and Policy Unit. A special thanks to Cecilie Bjørnskov-Johansen for helping me in balancing academia and practical development work. After Christian Friis Bach's initial encouragement, Allan Duelund Jensen also supported the work and finally, Birgitte Qvist-Sørensen was a supportive and committed industrial advisor in the last half of the project.

Also thanks to the Department of Business and Economics at the University of Southern Denmark for a welcoming and open-minded research environment. Jonas Helth Lønborg, whom I mentioned in the introduction, should be thanked not just for initiating the idea, but also for a large role in its execution all the way from field to findings. I am particularly grateful to Lene Holbæk, who has been kind to proof-read a large part of this thesis. Thomas Skovgaard Pedersen, a friend outside academia, provided great help on this as well.

During my PhD, I got the opportunity to re-visit Professor Michael Carter at the University of California-Davis. Michael initially got me interested in (development) economics, and he helped me to understand the possibilities and limitations of randomized control trials in economics. Along the way, I also benefitted from Christopher Ksoll's thorough experience in theory and practice of field research as well as Helene Bie Lilleør's support as something like an additional

advisor. The Rockwool Foundation and the Danish Agency for Science, Technology and Innovation funded the work.

The final words of thanks must and should go to my family, in particular my wife Stine. Not only did you encourage me to pursue something that you knew would take up a significant part of my time and mind and that put significant demands on our wonderful family. You also showed genuine interest in, and lent your brains to, the often peculiar issues that interested me along the way. I enjoy travelling with you.

INDLEDNING OG RESUMÉ PÅ DANSK (INTRODUCTION AND SUMMARY IN DANISH)

Virker udviklingsbistanden? Det spørgsmål optager både den brede befolkning, forskere og alle, der er involveret i arbejdet med international bistand. Og med god grund. Befolkningen har skattekroner på spil, og bistandsbranchen vil gerne gøre en forskel. Men spørgsmålet har også vist sig overordentligt vanskeligt at besvare, særligt hvis man stiller det på globalt niveau (Mankiw *et al.* 1995). På det lokale niveau er det lettere. Denne afhandling svarer derfor på spørgsmålet ved at undersøge effekten af et enkelt projekt i det nordlige Malawi, helt konkret etableringen af 100 spare-låne-grupper, som giver medlemmerne mulighed for at spare op og låne små beløb. Det kaldes også landsbybaseret mikrofinans.

Konklusionen i afhandlingens første kapitel er, at der er grund til forsigtig optimisme. Deltagerne i projektet oplever forbedret adgang til mad, højere generel levestandard og får mulighed for at udvide deres boliger, når man sammenligner dem med en kontrolgruppe. Undersøgelsen anvender en metode, der er udbredt inden for medicin: Ved hjælp af lodtrækning delte vi 46 landsbyer i to grupper i 2009. I halvdelen af landsbyerne tilbød en lokal organisation beboerne grundig uddannelse i organisering af låne-spare-grupper, herunder opsparing, indlån og regnskabsføring. Den anden halvdel af landsbyerne fungerede som kontrolgruppe, mens forsøget stod på, men fik dog efterfølgende samme tilbud. Vi måler effekten ved at sammenligne ændringerne på husholdningsniveau i de to grupper af landsbyer. Informationen i analysen stammer fra en spørgeskemaundersøgelse blandt 1.775 husholdninger, der er indsamlet med hjælp fra 24 lokalansatte og finansieret af Rockwool Fonden. Denne effektevaluering er fokus i kapitel 1.

De efterfølgende kapitler belyser en række relaterede spørgsmål: Deltager de fattigste i spare-låne-grupperne? Hvad er renten på opsparing? Og hvordan kan vi udvikle mere troværdige effektmålinger? Konklusionerne fra disse kapitler er, at de fattige deltager, dog ikke de allerfattigste. At grupperne giver medlemmerne

62 procent i årlig rente på opsparing, fordi de også låner pengene ud. Og at etableringen af et centralt globalt register til præregistrering empiriske studier vil kunne skabe større troværdighed omkring effektmålinger. De to sidste kapitler er udgivet i henholdsvis *Enterprise Development and Microfinance* (årgang 23, nummer 4) og *Journal of Development Effectiveness* (årgang 3, nummer 4).

I denne indledning vil jeg diskutere afhandlingens fokus og konklusioner i et bredere bistandsfagligt perspektiv og efterfølgende give en grundigere opsummering af de enkelte kapitler.

Lokale strategier i bistandsarbejdet

De bedst kendte strategier i international bistand er globale. Den nu 20 år gamle Agenda 21 danner rammen for bæredygtig udvikling. Den nyere Parisdeklaration har sat nye standarder for effektivisering af bistanden, og endelig beskriver millenniumdeklarationen fra 2000 de velkendte 2015-mål, der i øjeblikket er under revision. Alle disse globale strategier har sat hver deres dagsorden og har virket som løftestang i forhold til finansiering og udmøntning af bistandsindsatser.

Sådanne tiltag kan uden tvivl være en hjælp for lavindkomstlandene, fordi verdenssamfundet bliver enigt om prioriteterne i bistanden, og for eksempel administrationen af bistanden dermed bliver lettere. Det globalt accepterede mål om at halvere fattigdommen bidrager til et skærpet fokus, når vi skal allokere begrænsede midler. Men der er ikke megen praktisk vejledning at finde i sådan en målsætning. Hvis globale strategier skal ændre noget, så skal de følges op af lokal handling. Det centrale spørgsmål bliver: Hvad foregår der helt konkret i hverdagen blandt verdens fattigste? Det er spørgsmålet, jeg søger at besvare i denne afhandling, ved at bruge spare-låne-grupperne som eksempel.

Der er situationer, hvor det er ligetil at formulere en lokal strategi. Regeringer ved godt, hvordan man bygger veje og skoler, og hvordan man sørger for rent vand. Men selv i disse tilfælde skal vejene vedligeholdes og brøndene holdes fri for forurening. Det vil sige, at der er brug for en lokal strategi. Eller tag et eksempel som

ejendomsret. Hvordan håndterer man de mange, og til tider overlappende, krav på jord, der er baseret på sædvane? Her er der brug for en lokal strategi

Spare-låne-grupper er et eksempel på en meget udbredt lokal strategi i disse år. Grupperne består af 15 til 20 medlemmer som mødes, typisk en gang om ugen, for at spare op i en fælles bankboks og låne penge af deres fælles kapital. Grupperne har ingen ansatte, og der tilføres ingen kapital udefra. Det eneste, grupperne modtager, er et grundigt undervisningsforløb i, hvordan de kan håndtere pengene, så de ikke forsvinder. Som regel er det en lokal organisation, der sætter grupperne i gang. Der findes mindst 80.000 grupper med mere end to millioner medlemmer på globalt plan, ifølge det største register på områder (SAVIX 2012). Andre sammentællinger når frem til tre gange så mange (VSL Associates 2013). Spare-låne-gruppe-modellen kaldes også for landsbybaseret mikrofinans eller blot sparegrupper (Allen and Panetta 2010).

Der er flere grunde til, at det er interessant at studere spare-låne-grupper. For det første er de meget udbredte. For det andet er de bæredygtige i den forstand, at de vedbliver med at fungere selv flere år efter at undervisningen har fundet sted (Anyango and Esipisu 2007). For det tredje er de meget simple og dermed lette at forstå, både i lokalsamfundene og blandt dem, der står for undervisningen. For det fjerde har de ikke været genstand for væsentlig opmærksomhed fra forskerhold, ulig en anden type landsbybaseret mikrofinans, de uformelle, roterende spare-låne-grupper (Besley *et al.* 1993, Besley *et al.* 1994, Ardener and Burman 1995, Bouman 1995). Endelig kan spare-låne-grupper bruges i mange sammenhænge. Ofte er det erklærede formål med at etablere grupperne at bidrage til en højere økonomisk levestandard og for eksempel give bedre adgang til mad. Men grupperne kan meget vel tænkes også at have andre effekter. På grund af deres demokratiske procedurer for valg og håndtering af finanser kan det tænkes, at de udbreder kendskabet til demokrati og god forvaltningsskik. Hvis de bruges som afsæt for undervisning, kan de også bidrage til for eksempel folkesundhed. I

tilfælde, hvor grupperne tager hånd om problemer i lokalsamfundet, kan de endvidere bidrage til, at svage befolkningsgrupper bliver hørt.

Emnet for denne afhandling er det overordnede formål: Spare-låne-gruppernes rolle i at bibringe øget levestandard for de fattigste. Begrundelse for dette emnevalg er gruppernes økonomiske funktioner, som jeg beskriver i det næste afsnit.

Spare-låne-grupper som mikrofinans

I 2010 besøgte jeg en spare-låne-gruppe i det nordlige Malawi. Jeg spurgte deltagerne, hvorfor de var medlem af gruppen. En ung kvinde gav et særligt interessant svar. Hun sagde: ”Jeg er medlem, fordi min bedstemor netop er gået bort.” Da jeg ikke var helt sikker på, hvad hun mente, bad jeg hende uddybe. ”Da min bedstemor levede, holdt hun altid øje med mine penge for mig og hjalp mig med at få dem til at slå til. Hendes hus var det eneste sted, hvor jeg kunne opbevare dem sikkert. Nu, hvor hun er væk, har jeg ingen steder til pengene.”

Der er mange grunde til at blive medlem af en spare-låne-gruppe, og citatet ovenfor illustrerer en af dem: Behovet for et sikkert sted at spare up. Kvinden brugte sin bedstemor som bank, selvom bedstemoren ikke udbetalte renter. Kvinden havde brug for et sted at spare op. Det behov er hun langt fra den eneste, der har. Da nogle af de største mikrofinansinstitutioner for nogle år siden begyndte at tage imod opsparing, viste det sig, at behovet var enormt. BancoSol fra Bolivia og Grameen Bank fra Bangladesh har således begge et større beløb som indlån end som udlån (BancoSol 2009, Grameen Bank 2010). Nogle praktikere taler endda om en opsparingsrevolution. Revolution er måske så meget sagt, men rigtigt er det, at mikroopsparing får mere opmærksomhed i dag, end det gjorde for 30 år siden, hvor Robert Vogel udpegede opsparing som ”den oversete halvdel” af finansydelser til brug i udviklingsarbejdet (Vogel 1984). Den udvidede opmærksomhed er imidlertid også rettet mod en lang række andre finansielle ydelser, der tilbydes verdens fattigste ud over lån og opsparing, for eksempel forsikring, mobiltelefonbetaling og realkredit. Så det, der i sandhed er nyskabende, er anerkendelsen af det faktum, at adgangen til finansielle ydelser betyder noget, også

blandt verdens fattigste – eller måske særligt blandt verdens fattigste, på grund af den uforudsigelighed, der kendetegner indkomsten hos denne gruppe.

Uforudsigelig indkomst er væsentlig, når det kommer til finansielle markeders funktion. Det faktum, at 1,3 milliarder mennesker lever for mindre end en dollar om dagen kunne lyde, som om der faktisk er tale om en stabil indtægt på en dollar hver eneste dag. Det er ingenlunde tilfældet. Indtægten går op og ned, ikke mindste på grund af vejret, uanset om det drejer sig om indtægt fra landbrug, fiskeri eller en lille virksomhed. Forbruget, derimod, kan ikke variere i samme grad. De fleste foretrækker at spise hver dag, og det er simpelthen umuligt at spise for eksempel en gang om ugen.

En væsentlig funktion for finansielle markeder, uanset hvor strømlinede de er, er at muliggøre en tidsmæssig adskillelse af indtægt og forbrug. Opsparing er indtægt i dag, forbrug i morgen. En gruppe af kvinder fra Malawi, som jeg interviewede i 2010, og som ikke havde adgang til en spare-låne-gruppe, fortalte om opsparingsmetoder, der ligger meget langt fra bankvirksomhed, som vi kender det. En af kvinderne gravede pengene ned i jorden. En anden havde en særlig lomme i sin *chitenje*, den almindelige påklædning for kvinder på landet i Malawi, som hun syede til, hver gang hun havde lagt penge i den. En tredje gav pengene til sin nabo. En globalt repræsentativ undersøgelse fra 2011 bekræfter billedet: Mindre end 15 procent af den voksne befolkning i Afrika syd for Sahara sparede op på en bankkonto, mens 35 procent sparede op ad uformelle kanaler (Demirguc-Kunt and Klapper 2012). Analyser af finansielle dagbøger har konkluderet, at de fattigste ofte er meget aktive, når det kommer til at håndtere de daglige finanser. Data fra Sydafrika, Bangladesh og Indien viste, at en almindelig husholdning benyttede sig af ni forskellige steder til opsparing og kredit (Collins *et al.* 2009).

Hvis de finansielle markeder ikke fungerer, bliver det svært at få enderne til at nå sammen, særligt når indtægterne er både lave og uforudsigelige. Koger man det ned, så er formålet med al mikrofinans at skabe mere velfungerende finansielle markeder med udgangspunkt i bæredygtige institutioner. De almindelige mikrofinansinstitutioner,

der i høj grad minder om banker, har haft svært ved at nå ud på landet, særligt i Afrika, hvor befolkningstætheden er lav. Spare-låne-grupper, derimod, kan fungere på landet og er ikke afhængige af afstanden til større byer.

Lodtrækningsforsøg i bistanden

Denne afhandling analyserer effekten af spare-låne-grupper ved hjælp af et såkaldt randomiseret kontrolleret forsøg også kendt som et lodtrækningsforsøg. Denne metode, hvor man tilfældigt fordeler en gruppe af personer, eller i vores tilfælde landsbyer, i en kontrolgruppe og en gruppe, der deltager i en intervention, er i de seneste år blevet mere og mere populær som evalueringsmetode i både bistandsverdenen og i forskningssammenhænge. Der er generel enighed om, at metoden er god til at pege på årsagssammenhænge og svare på, om det var projektet eller andre forhold, der førte til ændringer i en given situation.

Imidlertid er der mange spørgsmål, som lodtrækningsforsøg ikke egner sig til at svare på. Det gælder tiltag som for eksempel reformer i den offentlige sektor eller budgetstøtte. Kritikere af metoden konkluderer derfor, at metoden kun bidrager meget lidt til at forstå bistanden (Deaton 2010). Omvendt mener fortalere for metoden, at tiltag, der ikke kan evalueres med lodtrækningsforsøg, for eksempel budgetstøtte, slet ikke bør gennemføres: "...vi bør gå tilbage til [kun] at finansiere projekter og insistere på, at resultaterne kan måles" (Banerjee 2007, p. 21, egen oversættelse).

Hvis dette synspunkt skulle føres ud i livet, ville det medføre dramatiske ændringer, ikke bare i måden, vi yder bistand på, men også i form af større administrative byrder, mindre grad af ejerskab hos målgruppen og ringere medbestemmelse for lavindkomstlandenes regeringer og befolkninger. Bistandsindsatser og bestemte måder at give bistand på kan være nyttige og brugbare, uanset om de kan udsættes for et lodtrækningsforsøg eller ej. Det tiltag, eller den politik, der evalueres, bør bestemme evalueringsmetoden – ikke den anden vej rundt.

Men der er situationer, hvor lodtrækningsforsøg kan anvendes og endda tilfælde, hvor de er særdeles velegnede til at måle, i hvilket omfang en bistandsform er effektiv. I disse tilfælde bør muligheden for troværdig viden om effekt ikke forbigås. Denne viden bør indgå i planlægningen af fremtidige indsatser, ikke mekanisk og uden hensyntagen til anden viden og lokale forhold, men som ligeværdig del af det foreliggende beslutningsgrundlag. Dermed kan den bidrage til læring blandt de personer og organisationer, der beskæftiger sig med bistand.

Låne-spare-grupper er kun en lokal strategi blandt mange, og jeg kommer kun ind på en lille del af de spørgsmål om grupperne, der er relevante. Blandt de spørgsmål, jeg ikke diskuterer, er de interne dynamikker i grupperne, forsikringens rolle, de ikke-økonomiske konsekvenser af at deltage samt konsekvenser for forholdene internt i husholdningerne. Jeg håber, at fremtidig forskning vil bidrage til at kaste lys over disse spørgsmål, og at de fire kapitler, som jeg opsummerer i det følgende, kan tjene som inspiration og udgangspunkt.

”Virker bistanden?” er et vanskeligt spørgsmål at stille, eftersom svaret muligvis er ”nej.” Det er endnu vanskeligere at besvare, selv når man fokuserer på noget så tilsyneladende simpelt som spare-låne-grupper. Ikke desto mindre stiller jeg spørgsmålet. Det gør jeg af tre grunde. For det første fordi dem, der betaler, hvilket ofte er skatteyderne, har krav på at vide, om deres penge bliver brugt fornuftigt. For det andet for at svaret kan bidrage til videreudviklingen af metoderne inden for international bistand. Og sidst, men på ingen måde mindst, af hensyn til de lokalbefolkninger, der bruger deres tid og kræfter på at deltage i projekter som det, der er beskrevet i det følgende.

Resumé af afhandlingen

Afhandlingen består af fire kapitler. De to første kapitler er baseret på data fra den tidligere omtalte spørgeskemaundersøgelse blandt 1.775 husholdninger, som blev udført i forbindelse med afhandlingen.

Analyserne i de to sidste kapitler bygger på data fra henholdsvis CARE Malawi og et offentligt tilgængeligt register.

Det første kapitel er skrevet med Christopher Ksoll, Helene Bie Lilleør og Jonas Helth Lønborg. Kapitlet omhandler, som allerede beskrevet, resultaterne af en effektmåling af et projekt, der i løbet af en toårig periode igangsatte 100 spare-låne-grupper i det nordlige Malawi. Ved at lade en lodtrækning afgøre, hvilke af de 46 landsbyer, der skal agere kontrolgruppe, sikrer vi, at kontrolgruppen er sammenlignelig med gruppen af landsbyer, der deltager i projektet.

Vi finder positive effekter på fire ud af ti på forhånd fastsatte indikatorer. For det første øges adgangen til mad målt som det antal måltider, en husholdning spiste dagen før undersøgelsen. Husholdningerne i projektlandsbyerne spiser i gennemsnit 0.13 måltider mere end kontrolgruppen, hvilket . For det andet stiger det generelle forbrug med tre procent og endelig udvider husholdningerne i projektlandsbyerne deres boliger med 0.16 værelse i gennemsnit. Da der er tale om et meget fattigt område, er disse ændringer betydningsfulde. Vi finder ingen effekter på længden af den såkaldte sultne periode lige før høsten, hvor folk spiser mindre end normalt og heller ikke på den samlede opsparing, jordbesiddelser, forbrug på 17 udvalgte fødevarer eller kvaliteten af boligen. Ud over effekten ser vi på, baggrunden for den positive udvikling og konkluderer, at en af årsagerne er øget investering i landbrug.

I det andet kapitel ser jeg, sammen med Jonas Helth Lønborg, nærmere på, hvorvidt de fattigste deltager i spare-låne-grupperne. Efter en gennemgang af litteraturen videreudvikler vi de eksisterende metoder til at afgøre, om de fattigste deltager. Resultaterne viser, at ganske vist er der mange fattige, der deltager, men de allerfattigste er underrepræsenterede i projektet. En nærmere analyse viser, at det kan skyldes, at den indledende informationskampagne ikke når tilstrækkelig mange og at de allerfattigste ikke kan honorere gruppens krav om opsparing.

I tredje kapitel undersøger jeg renten på opsparing i grupperne. Når grupperne låner de opsparede midler ud, genererer de rente, som bliver fordelt blandt medlemmerne. Med data fra 204 grupper i det

centrale Malawi finder jeg, at den årlige opsparingsrente er 62 procent, hvilket er omtrent det dobbelte af, hvad der er den almindelige opfattelse (for eksempel hos Allen and Panetta 2010). Forskellen opstår, fordi praksis på området ikke følger finansiell standard for rentesregning, og jeg udvikler en række metoder, der gør det nemmere at anvende den korrekte udregning i praksis. Kapitlet er udgivet i *Enterprise Development and Microfinance*, 23. årgang, nummer 4.

Det sidste kapitel, som er skrevet sammen med Thomas Barnebeck Andersen og Nikolaj Malchow-Møller, er både en kritik af kvantitativ empirisk metode, særligt lodtrækningsforsøg, samt et praktisk forslag til, hvordan metoden kan forbedres. Et problem i mange kvantitative empiriske undersøgelser, heriblandt lodtrækningsforsøg som det, der er beskrevet i kapitel 1, er, at de forholder sig til mange variable på én gang. Dermed er der en forhøjet risiko for falske positiver: effekter, der ikke er sande, men som man finder ved ren tilfældighed. Et register, hvor man, inden man påbegynder et studie, kan registrere, hvilke variable man vil måle på, vil bidrage til at afhjælpe problemet. I kapitlet gennemgår vi erfaringerne fra sådanne registre inden for medicin og kommer med forslag til, hvordan man kan etablere et specifikt register for empiriske studier i lavindkomstlande. Dette kapitel er udgivet i *Journal of Development Effectiveness*, 3. årgang, nummer 4.

Referencer

- Allen, H. and Panetta, D. 2010. *Savings Groups: What Are They?* Washington DC: The SEEP Network.
- Anyango, E. and Esipisu, E. 2007. Village Savings and Loan Associations - experience from Zanzibar. *Small Enterprise Development*, 18(1), pp. 16.
- Ardener, S. and Burman, S. (eds.) (1995) *Money-Go-Rounds*, Oxford: Berg.
- BancoSol 2009. *Annual Report 2008*. La Paz: BancoSol.
- Banerjee, A. 2007. *Making aid work*. Cambridge, Mass.: MIT press.

- Besley, T., Coate, S. and Loury, G. 1993. The economics of rotating savings and credit associations. *The American Economic Review*, 83(4), pp. 792-810.
- Besley, T., Coate, S. and Loury, G. 1994. Rotating savings and credit associations, credit markets and efficiency. *The Review of Economic Studies*, 61(4), pp. 701-719.
- Bouman, F. 1995. Rotating and accumulating savings and credit associations: A development perspective. *World Development*, 23(3), pp. 371-384.
- Collins, D., Morduch, J., Rutherford, S. and Ruthven, O. 2009. *Portfolios of the Poor: How the World's Poor Live on \$2 a Day*. Princeton, NJ: Princeton University Press.
- Deaton, A. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2), pp. 424-455.
- Demirguc-Kunt, A. and Klapper, L. 2012. *The Global Findex Database. New Data on Saving, Borrowing, and Managing Risk*. Findex Note. Washington D.C.: World Bank.
- Grameen Bank 2010. *Grameen Bank At A Glance*. Dhaka: Grameen Bank website, available at <http://grameen-info.org> [Accessed May 2010].
- Mankiw, N.G., Phelps, E.S. and Romer, P.M. 1995. The growth of nations. *Brookings papers on economic activity*, 1995(1), pp. 275-326.
- SAVIX 2012. *Savings Groups Information Exchange*. Germany: VSL Associates. Available from <http://savingsgroups.com>. [Accessed January 2013].
- Vogel, R.C. 1984. Savings mobilization: The Forgotten Half of Rural Finance. In D. Adams, D. Graham and J.D. Von Pischke (eds) *Undermining Rural Development with Cheap Credit*. Boulder, CO, USA: Westview Press.
- VSL Associates 2013. *Global Outreach of Savings Groups: Current information*. Solingen, Germany: VSL Associates. Available at

<http://vsla.net/outrich/INGO-CMMF-Global-Outreach.2012-06.xls> [Accessed May 2013].

INTRODUCTION AND SUMMARY OF CHAPTERS

Does aid work? That is one of the most frequently asked questions by practitioners and academics in international development. On a global scale, the question has proved difficult to answer (Mankiw *et al.* 1995). Locally, it is easier. So in this thesis, I provide at least part of the answer by evaluating a project that has established 100 savings and loan associations – small groups that provide financial services to members – in northern Malawi. In this particular case, the answer seems to be yes: the first chapter demonstrates that savings and loan associations have positive effects, particularly on food security. The method is borrowed from medicine: a total of 46 villages are randomly divided into two groups, and in half of the villages, savings and loan associations are established, whereas in the other half, business as usual prevails. In total, 1,775 households are surveyed. This work was carried out by a local staff of twenty-four interviewers and supported by the Rockwool Foundation.

The impact of savings and loan associations is at the core of the thesis. The remaining three chapters investigate a number of related issues: Who participates? What is the interest rate in the groups? What would it take to increase the credibility of results? The findings are that the poor participates, but the poorest less so. Groups generate a sixty percent interest rate on savings. And the establishment of a trial registry could improve credibility of results in the future. The two last chapters have been published in *Enterprise Development and Microfinance* (vol. 23, iss. 4) and *Journal of Development Effectiveness* (vol. 3, iss. 4), respectively.

In the following sections I will place the study in a broader discussion on aid, introduce savings and loan associations and provide summaries of the four chapters.

In Defense of a Local Strategy

There is no shortage of global strategies in international development. Examples are Agenda 21 on sustainable development, the Paris Declaration on aid effectiveness or, perhaps the most well-known, the

millennium declaration, which contains the eight millennium development goals (MDGs), and is currently under revision. All of these have shaped the agenda on aid in the past decades and have created a common ground for funding and implementation of programmes.

Aligning priorities globally can reduce the administrative burden of aid experienced by developing countries. Accepting the goal of halving global poverty, for example, makes us focus our attention and resources. But it does little in terms of guiding practical action. To be effective, every global strategy needs local action. We need to continue asking: What happens on the ground? Using the example of savings and loan associations, that is the overall question asked in this thesis.

Sometimes the local strategy is straightforward. Governments know how to build roads, provide water or ensure access to energy. But even then, the roads need to be maintained, and wells must be free of contamination. A local strategy is needed. In other cases it is much harder. Take the promotion of property rights, something often said to foster both social progress and economic growth. How do you manage the many customary, but sometimes overlapping, claims people rightfully have on their land? Certainly, you need a local strategy. A final example closer to home: many civil society organizations base their work on human rights and seek to empower people and build capacity with local government. Success clearly depends on intimate understanding of the local context and a clear idea of what happens on the ground.

The concept of savings and loan associations (SLAs) is an example of a very commonly used local strategy. SLAs are groups of 15 to 25 people who get together to save and borrow among each other. There is no staff employed by the groups and no outside provision of capital, but merely a series of training sessions and thorough supervision. Usually initiated by a local organization, there are now at least 80,000 groups with more than 2m members if we count only members of groups registered at <http://savingsgroups.com>. Other counts find three times as many (VSL Associates 2013). The SLA model is also known

as community-managed microfinance, village savings and loan associations or, simply, savings groups (Allen and Panetta 2010).¹

There are several reasons why SLAs are worth studying. First, the method is widely used as the figures mentioned above show. Second, they seem to be very sustainable. After a year of training groups have been found to exist and proliferate many years later (Anyango and Esipisu 2007). Third, they are easy to understand by communities and implementers due to their simplicity. Fourth, they have not been thoroughly studied elsewhere. Their cousin and predecessor, the rotating savings and loan associations, have been subject to much academic attention (Besley *et al.* 1993, Besley *et al.* 1994, Ardener and Burman 1995, Bouman 1995). With the SLAs, less so, even though it is on the rise. Finally, SLAs can potentially be used as a local strategy for many purposes. There is the direct and stated purpose of promoting food security and economic welfare. But there are other possible effects, too. Due to their democratic structure, the groups might serve as an exercise in democracy and accountability. As a platform for training, they can be used in public health programming. And when groups take on community issues, they form the basis of local advocacy work.

The topic of this thesis is primarily the first one: The role in SLAs in promoting food security and economic welfare. To motivate this, the next section discusses the economic role of SLAs.

SLAs as Microfinance

In 2010 I visited a savings and loan association in northern Malawi. I asked the participants why they were members of the group. Among

¹ The reason I call them savings and loan associations and not, for example, savings groups, is that I want to emphasise two aspects: That groups provide loans as much as savings. And that they are associations in the sense that they have a constitution, a board and elections. This is in line with Pors (2011). Nevertheless, since the separate chapters have slightly separate audiences, I use different terms in different chapters. “Village savings and loan associations” is used in chapter one, since this is the particular SLA methodology we study there. In chapter two we use “community-management microfinance,” a broader term, to emphasize the connection to other types of microfinance. Finally, “savings groups” is used in chapter three, since this is the term used mostly by practitioners.

the different answers, one young woman's answer caught my attention. She said, "I am a member because my grandmother just passed away." Not completely sure how to interpret this, I asked her what she meant. "When my grandmother was alive, she looked after my money and helped me to organize my finances. She was the only place I had to store money. Now that she's gone, there is no place left."

People join these groups for a variety of reasons, and the quotes above illustrate one of these reasons: the need for a safe place to save. The woman used her grandmother's deposit service, even though it earned her no interest. She had a clear demand for a savings service. Recent examples of microfinance institutions which have started to accept deposits from poor clients have demonstrated a huge demand for savings services. BancoSol and Grameen Bank, two leading microfinance institutions, both have more deposits than credit (BancoSol 2009, Grameen Bank 2010). Some practitioners are even talking about a "savings revolution". Revolution or not, it is certainly the case that microsavings today gets more attention than it did 30 years ago when Robert Vogel called savings "the forgotten half of rural finance" (Vogel 1984). But so do other financial services provided to poor populations, like insurance, mobile money transfers or mortgages. So the real change is the recognition of the fact that financial management is important, also among the world's poorest - or perhaps in particular among the world's poorest, due to the unpredictability of income for this group.

Unpredictable income is central to understanding the role of financial markets. The fact that 1.3bn people live on less than one dollar per day may sound as if they somehow receive one dollar every day. Nothing could be further from the truth. Income varies tremendously, whether it is in-kind income from farming, proceeds from selling fish or profits from a small business. Consumption, on the other hand, cannot vary too much. Most people prefer to eat every day, and eating only one time per week is simply impossible.

Financial markets, no matter how unorthodox, often work to enable separation of income and consumption in very practical ways:

consume tomorrow what you earn today. A group of women in Malawi, who did not have access to a savings and loan association, revealed savings practices very far from those of Wall Street financial markets, during a focus group interview I did in 2010. One buried the money in a pot in the ground. Another had a special pocket in her *chitenje*, the most common women's wear of Malawi, which she stitched back together after each deposit. Yet another deposited money with her neighbour. A global representative survey from 2011 confirms this picture: less than 15% of the adult population in sub-Saharan Africa saved in a formal account, whereas 35% saved using informal means (Demirguc-Kunt and Klapper 2012). The use of so-called financial diaries has shown that the poor are often very active when it comes to managing finances. Evidence from South Africa, India and Bangladesh shows that the average household have as many as nine different places to save and take loans (Collins *et al.* 2009).

If financial markets do not work, it is difficult to make ends meet when income is low and uncertain at the same time. In essence, the purpose of microfinance is to make these markets work better, while still building sustainable institutions. More bank-like microfinance, which includes almost all of microfinance in the developing world, has had a hard time reaching rural Africa, probably due to the high transaction costs, primarily transport. SLAs, on the other hand, do not depend on outside contact once they are established.

Making Sense of Randomization in Aid

This thesis provides evidence of effect of savings and loan associations using a randomized control trial (RCT). The method of randomly assigning people or villages to interventions and control groups has gained increased popularity, not the least in development aid and development economics. There is general agreement that an RCT is a good way of identifying a causal link, something which is often challenging, but the method has nevertheless spurred its own debate. Critics argue that RCTs are not suitable when it comes to evaluating a wide range of common development interventions, for example public sector reform or budget support (Deaton 2010). That

randomization is only possible in some cases, is a matter of fact. But the consequence of this is not. Some RCT proponents draw the logical conclusion and propose that “we need to go back to financing projects and insist that the results be measured” (Banerjee 2007, p. 21). If followed through, this would have wide ranging consequences, not just for the global aid architecture, but also for the administrative burden, ownership and level of participation for developing world governments and people. Policies and aid modalities can be necessary and useful, even though they cannot be randomized. The policy being evaluated should decide the method, not the other way around. But sometime questions can in fact be answered using RCTs, and in more than a few cases, RCTs are even very good at it. In those cases, the possibility of better evidence should not be foregone. This evidence should enter the planning cycle, not in a mechanical way that disregards all other evidence or non-randomizable policies, but as one piece of qualified input to a discussion among learning professionals in organisations working in with development aid.

The concept of savings and loan associations is just one local strategy among many, and I cover only a small part of the relevant issues regarding this strategy below. Some questions I do not explore are the internal dynamics of the groups, the role of insurance, the non-economic consequences of joining and intra-household effects of group membership. I hope that future research will address some of these issues and that the four chapters I summarise below will provide a useful starting point.²

“Does aid work” is a tough question to ask, since the answer could be “no.” It is even harder to answer, even when evaluating something as simple as savings and loan associations. But I ask it nevertheless for at least three different reasons: to ensure that public funds are well spent, to improve the methods of international development aid, and, last but certainly not least, for the sake of local communities, who

² The introduction and the first two chapters are written in American English. The last two chapters use British English. This inconsistency is due to the language policies of the journals where the chapters have been or will be submitted.

often spend significant time and effort participating in interventions like the one described in the following.

Summary of Chapters

Chapter 1: The Impact of Village Savings and Loan Associations on Household Welfare: Evidence from a Cluster Randomized Trial³

This chapter reports the main findings from an impact evaluation of savings and loan associations in 46 villages in northern Malawi. Prior to the introduction of the savings and loan associations, the 46 villages were divided randomly into two groups. In the first group, the so-called treatment group, savings and loan associations were introduced in 2009, whereas in the second group, the control villages, implementation was delayed until 2011. A random sample with 1775 households in all 46 villages was surveyed just before the first groups were established in 2009 and again before introducing the groups in control villages in 2011. The random division into treatment and control groups ensures that we should expect no systematic differences between the two groups besides those that are due to the introduction of savings and loan groups.

We find positive effects on a number of the pre-specified outcomes, the most robust being food consumption as measured by number of meals consumed per day. We also find indications of an increase in household income, a decrease in the length of the hungry period, an increase in total savings and an increase in size of house. We do not find any effects on the number of income-generating activities, size of land used for agriculture or housing quality.

After presenting the main results, we investigate possible channels of impact. Perhaps surprisingly, we find that the most probable channel is increased investment in agriculture. Once a year, the funds in the savings and loan associations are distributed among the members, and participants report that they use this to invest in agricultural inputs. We find support of this in the data. Households in treatment villages use more fertiliser and invest more in improved

³ Joint work with Christopher Ksoll, Helene Bie Lilleør and Jonas Helth Lønborg.

seed varieties that respond particularly well to fertilizer use. The results are promising in terms of savings and loan associations in general, but particularly regarding agriculture, where savings and loan associations can serve as an alternative to other, often more expensive, interventions aimed at increasing investment in agriculture, for example Malawi's fertiliser subsidy programme.

One methodological innovation is that the study evaluates the project according to a set of original project outcomes. If we had investigated the effects on all possible outcomes, we would most likely have found at least some significant effects, simply by chance. This is known as "data mining". To reduce the chance of this happening, we measure impact only on the outcomes specified in the project's documents, which were determined in advance.

A caveat in the study is that we find significant differences on several outcomes in our baseline survey when comparing treatment and control villages. This is a problem since the effect we find can then be driven by this initial difference instead of being the effect of the intervention. We do two things to argue that this is not the case. In the analysis, we make use of the baseline data, so we analyse the difference in developments in the treatment and control groups, and not the levels. We also show that the pre-program trends are similar in the two groups.

Chapter 2: Can Microfinance Reach the Poorest? Evidence from a Community-managed Microfinance Intervention⁴

An explicit goal of many development interventions is to reach the poorest, and microfinance as such is no exception. It is often recognized, however, that some people cannot make use of particularly credit-driven microfinance. This schism, that microfinance aims to reach the poorest but probably cannot, is the topic of chapter two. SLA methodologies were developed specifically with the purpose of reaching the poorest and therefore make a good case to examine the extent to which it is possible to reach the poorest.

⁴ Joint work with Jonas Helth Lønborg.

To do that, we single out a useful performance measure in the literature, which we call the outreach ratio. This is calculated as the share of poor people participating divided by the share of poor people in the population in general. The advantage of this metric is that it allows comparison across interventions and variables. If this ratio is above one, targeting is progressive and the intervention has successfully targeted people poorer than average. If the ratio is below one, targeting is regressive, which is to say that the intervention reaches people richer than the average. In the past, the outreach ratio has been calculated on the basis of various poverty metrics. After reviewing the literature, we supplement these outreach ratios with one we develop based on the so-called squared income gap from the poverty measurement literature. This metric is particularly useful when focusing on the poorest of the poor.

We find clear signs of regressive targeting and more so when using our newly developed metric. The intervention does reach a large number of poor people, but the average in the area is even poorer, a result which is robust to different poverty metrics. When asking the people why they do not join, some say that they do not have money to save, a finding which contradicts common knowledge in the field.

Finally, we investigate participation sequentially by looking at the so-called participation pipeline. Exiting is possible through various “holes in the pipeline”: not receiving information about the intervention, not being interested in joining, not finding a group to join, dropping out of a group and not participating fully in the group’s activities. We draw three conclusions. First, more than one third of the population in the area did not hear about the intervention through the awareness campaign. Second, of those who did, almost all were interested, but many did not join despite being interested. Third, those who join at this stage are poorer than those who do not. We conclude that the awareness campaign did not adequately explain the intervention.

Chapter 3: Small Groups – Large Profits. Calculating Interest Rates in Community-Managed Microfinance⁵

The interest rates in SLAs are important because profitable lending, and thus the creation of interest, is a prerequisite for effect. Moreover, interest rates matter when we want to compare SLAs to other financial provides.

In that regard, it is common in SLA practice to hear the following quip: “Interest rates are 10% per month on loans and 30% annually on savings.” The quote has appeared in *The Economist* as well as on the global webpage of savingsgroups.com. There are two issues with this statement, and chapter three seeks to address both of them. First, an attentive observer will see that the sentence uses two different time periods, making the two interest rates incomparable. That is easy to deal with but reveals a stark difference between the interest rate on savings and that on loans, since the interest rate on loans becomes 245% (due to 13 four-week periods in a year). Second, it turns out that the interest rate on savings is calculated using non-standard interest rate calculation, and this makes the two even more incomparable.

To assess the interest rate, I use data from 204 SLAs in Malawi to show that the global interest rate on savings is likely to be at least 60%, not 30. Indeed, in the 204 groups I look at, the median interest rate on savings using standard financial calculation is 62%, while the commonly used simple calculation gives 29%. I can employ the financial calculation because I have four observations per group during their first year. The simple calculation is wrong by more than 18 percentage points in 75% of the groups. The primary reason for this discrepancy is that the simple calculation assumes that all funds are saved at the beginning of the period, whereas in fact it is linear on the average.

On the practical side of things, I suggest a simple way of calculating the interest rate in savings groups using available data, without having information about the savings profile over time. Specifically, I develop a look-up table which translates the simple

⁵ This chapter is published in 2012 in *Enterprise Development and Microfinance*, 23(4), pp. 298-318.

interest rate into the financial interest rates and show that this is, on average, just as good as the financial calculation.

Even with the new calculation, there is still a large difference between the interest rates on savings and loans, but most of this can be explained by the fact that not all the funds are lent out all the time. I end the paper by suggesting a method that takes this into account in order to identify the “missing interest rate”, which is the difference between the return on savings a group should have made and the return it actually makes. I argue that this is a good metric to use when monitoring groups.

Chapter 4: Walking the Talk: the Need for a Trial Registry for Development Interventions⁶

Impact evaluations that use random allocation of people or groups into a control group and a treatment group have gained widespread popularity in economics in general and in development economics in particular. The reason is simple: when the two groups are selected at random, there are no systematic differences between them. Any difference must be a result of the intervention being evaluated. Advocates believe that evaluations using this method have the highest credibility, or internal validity, possible. When the assignment is random, a correlation between treatment and outcome is indeed causation.

Chapter four questions this assertion by focusing on one simple threat to the link between correlation and causation: randomness. Randomized control trials in development economics usually measure outcomes using surveys with literally hundreds of questions, giving researchers the same number of variables to analyse. In a completely random dataset, there will be positive significant effects on some of these variables that are created by chance. It is easy to find such random results, a practice termed “data mining”. In the medical field this is solved by forcing researchers to commit to a small number of outcome variables prior to undertaking a randomized trial and

⁶ The chapter is published in 2011 in *Journal of Development Effectiveness*, 3(4), pp. 502-519. Joint work with Nikolaj Malchow-Møller and Thomas Barnebeck Andersen.

registering these variables in a trial registry. We suggest the same solution for development economics.

A trial registry would not only reduce the chance of false positives created by sheer chance, but would also provide some additional benefits. First, registration of all studies would provide the full picture of all initiated studies, including those which in the end show no effects and thus might not be published due to publication bias. Second, it would be easier for researchers to credibly investigate effects that are believed to be unlikely by the majority of their peers. Today, such non-standard effects are likely to be discarded as data mining, but with a trial registry, researchers would be able to pre-commit to these outcomes.

To illustrate what it takes to establish a trial registry, we review the experience from the last 50 years in the medical field. We find that, apart from legislation, one single event seems to drive registration: the decision in September 2005 by 850 journals editors that studies must be registered to be published. The average daily registrations tripled. On this basis, the chapter ends with the suggestion that journal editors join standard-setting organizations like the OECD to promote a trial registry for development interventions.

References

- Allen, H. and Panetta, D. 2010. *Savings Groups: What Are They?* Washington DC: The SEEP Network.
- Anyango, E. and Esipisu, E. 2007. Village Savings and Loan Associations - experience from Zanzibar. *Small Enterprise Development*, 18(1), pp. 16.
- Ardener, S. and Burman, S. (eds.) (1995) *Money-Go-Rounds*, Oxford: Berg.
- BancoSol 2009. *Annual Report 2008*. La Paz: BancoSol.
- Banerjee, A. 2007. *Making aid work*. Cambridge, Mass.: MIT press.
- Besley, T., Coate, S. and Loury, G. 1993. The economics of rotating savings and credit associations. *The American Economic Review*, 83(4), pp. 792-810.

- Besley, T., Coate, S. and Loury, G. 1994. Rotating savings and credit associations, credit markets and efficiency. *The Review of Economic Studies*, 61(4), pp. 701-719.
- Bouman, F. 1995. Rotating and accumulating savings and credit associations: A development perspective. *World Development*, 23(3), pp. 371-384.
- Collins, D., Morduch, J., Rutherford, S. and Ruthven, O. 2009. *Portfolios of the Poor: How the World's Poor Live on \$2 a Day*. Princeton, NJ: Princeton University Press.
- Deaton, A. 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2), pp. 424-455.
- Demirguc-Kunt, A. and Klapper, L. 2012. *The Global Findex Database. New Data on Saving, Borrowing, and Managing Risk*. Findex Note. Washington D.C.: World Bank.
- Grameen Bank 2010. *Grameen Bank At A Glance*. Dhaka: Grameen Bank website, available at <http://grameen-info.org> [Accessed May 2010].
- Mankiw, N.G., Phelps, E.S. and Romer, P.M. 1995. The growth of nations. *Brookings papers on economic activity*, 1995(1), pp. 275-326.
- Pors, K.K. 2011. *Savings- and loan associations. Combining short term benefits with long term structural change*. Copenhagen, Denmark: Danish Forum for Microfinance and Danish Mission Council Development Department.
- Vogel, R.C. 1984. Savings mobilization: The Forgotten Half of Rural Finance. In D. Adams, D. Graham and J.D. Von Pischke (eds) *Undermining Rural Development with Cheap Credit*. Boulder, CO, USA: Westview Press.
- VSL Associates 2013. *Global Outreach of Savings Groups: Current information*. Solingen, Germany: VSL Associates. Available at <http://vsla.net/outreach/INGO-CMMF-Global-Outreach.2012-06.xls> [Accessed May 2013].

Chapter 1

The Impact of Village
Savings and Loan
Associations on Household
Welfare: Evidence from a
Cluster Randomized Trial

The Impact of Village Savings and Loan Associations on Household Welfare: Evidence from a Cluster Randomized Trial

Christopher Ksoll
University of Ottawa

Helene Bie Lilleør
Rockwool Foundation Research Unit

Jonas Helth Lønborg
University of Southern Denmark

Ole Dahl Rasmussen
University of Southern Denmark and DanChurchAid

Abstract: Seventy percent of the world's poorest live in rural areas with poor access to finance. Improving their access to credit and savings has the potential for facilitating investments, improving risk-coping, and enabling consumption smoothing. Using a cluster randomized trial, we investigate the impacts of a community-managed microfinance intervention aiming to do exactly this. We assess the impact of village savings and loans associations (VSLAs) in forty-six villages from Northern Malawi. Based on a panel survey of 1,700 households from the treatment and control villages, we find that the intervention had positive and significant effects on the number of meals consumed in a day, total household consumption, and number of rooms in the dwelling. This effect is linked to an increase in the use of fertilizer, increasing the maize agricultural production in the treatment villages. These results are robust to using first-differencing and adjusted regressions.

Acknowledgements: The authors wish to thank the implementing partners DanChurchAid and Livingstonia Synod Development Department (SOLDEV), in particular the field team led by Waluza Munthali, the Invest in Knowledge Initiative for data collection, and the Rockwool Foundation for financial support. We also thank participants at the CSAE, NOVAFRICA, UC Davis Agricultural and Applied Economics and DGPE workshops as well as Thomas Barnebeck Andersen, Nikolaj Malchow-Møller, Måns Söderbom, Peter Sandholdt Jensen, Henrik Hansen, and David McKenzie for useful comments.

1. Introduction

More than seventy percent of the world's 1.3 billion extreme poor live in rural areas. Rural economies are characterized by long time spans between input and output of the agricultural production, high degree of uncertainty in output prices, and little influence over the key input: rain. All of these characteristics create a high demand for financial services to enable investment, consumption smoothing, and risk coping (Conning and Udry 2007). Nevertheless, the history of rural financial intermediation is not encouraging. Even the recent explosive growth in microfinance globally has concentrated in urban and semi-urban areas (Daley-Harris 2009, Allen and Panetta 2010, Demirgüç-Kunt and Klapper 2012).

When formal institutions are not available, households use informal savings, credit, and insurance mechanisms instead. The widespread use of ROSCAs, ASCAs, susu-collectors, and similar informal financial networks is a testament to this fact (Aryeetey and Steel 1994, Rutherford 2001, Aryeetey 2005, Collins *et al.* 2009). One intervention which has gained increased popularity in rural Africa is savings and loan associations, or savings groups, where fifteen to twenty-five members are taught how to manage their own savings and borrowing. Savings groups provide a higher degree of flexibility, transparency, and security than most informal alternatives.

At least 60,000 such groups were known to exist in 2009, with membership typically between fifteen and twenty-five members (Allen and Panetta 2010). Despite their prevalence and despite the amount of donor funds used for starting up these groups by, for instance, CARE, thorough impact assessments are only recently starting to appear.⁷ This is unlike the situation in more bank-like microfinance, where several impact studies exist and show limited effects. See Stewart *et al.* (2010) and Copestake *et al.* (2011) for reviews of the existing evidence.

⁷ We know of a number of consultancy reports, of which two very recent ones use randomized control trials. The only research paper we know of cannot be cited at the time of writing, unfortunately.

The purpose of this paper is to provide an assessment of the impacts of one highly standardized type of savings groups, Village Savings and Loan Associations (VSLAs), on household outcomes. We do this through a cluster randomized control trials in forty-six villages in northern Malawi. Among these forty-six villages, twenty-three were randomly selected to participate in a VSLA project implemented by a local NGO from 2009 onwards. The remaining twenty-three villages were used as a control group in this setup, before the project was finally extended to these in 2011. We conducted household surveys of 1775 households in 2009, just before the project was implemented in the treatment villages, and again in 2011, before the intervention was introduced in the control villages. We carefully tracked households that had moved in between the time of the 2009 and 2011 surveys, resulting in a low attrition rate of only three percent. In the primary analysis, we assess the impact of the intervention on three groups of primary outcomes: food security, income-generating activities, and household income. These outcomes were defined by the project team in advance, and thus, although they were not pre-specified down to the individual variables, we do believe the opportunity was reduced for conscious or unconscious search for significant effects through data mining (Rasmussen *et al.* 2011, Casey *et al.* 2012).

Estimating intention-to-treat effects (ITT)—defined as the average effect among households of belonging to a village that was assigned to treatment—we find positive and significant effects on some outcomes and no effect on others. Among the food security outcomes, the number of meals eaten the day before the survey increased by 0.13 meals. We found no effect on the length of the hungry period or on food consumption. Among the three outcomes for income-generating activities, there are no significant impacts on the number of these activities or on total savings, but there are positive effects on the amount of savings in the VSLA, indicating that the intervention did mobilize savings. In the final group of outcomes, we find two significant effects: An increase in the estimated total consumption

using USAIDs Poverty Assessment Tool (PAT) methodology⁸ of 3.2% per day. And an increase in the number of rooms in the dwelling by 0.15, while there is no significant effect on the type of floor in the dwelling or the amount of agricultural land owned.

Apart from the analysis on the outcomes predefined by the project team, we also look at a number of intermediate outcomes to shed some light on the potential channels through which the intervention could have had this observed effect. Beyond the increase in savings in VSLA, we find significant increases in the use of credit, including credit used for investment purposes, according to participants' own statements. Likewise, respondents report that they use their savings for investments, primarily in agriculture.⁹ We pursued this flow of money into agriculture and small scale businesses, finding a statistically significant increase in the use of fertilizer and irrigation, followed by an increase in the value of the maize harvest. Several other agricultural outcome variables showed no effect, and there was no significant increase the total income from enterprises, although we did signs of more enterprises starting. On the cost side, we conduct a very simple cost-benefit analysis. Assuming that effects last for one year, the analysis shows that the intervention generates additional consumption worth USD 58\$ and costs USD 35\$ per household in treatment villages.

We conclude that overall this intervention had small, but statistically significant and positive, effects over the two-year implementation period assessed here. This is despite problems with households not complying with the randomization, resulting in an only twenty-three percentage point difference in the take-up of the intervention across treatment and control villages, just as only sixteen percent of the households had distributed savings back to the group members in a so-called share-out, which happens once per year. Whereas this result is promising for savings groups as a whole, it is

⁸ See <http://www.povertytools.org> for further information about the Poverty Assessment Tools.

⁹ The savings, along with interests earned, are distributed to each group member at the time of the yearly so-called "share-out," which will be described in further detail below.

important to remember that our analysis covers only one geographic location at one particular point in time. More evaluations are needed to judge whether this result can be generalized and to assess long-term effect.

This paper is structured as follows. In the following section, we describe the intervention and the research design. Section three explains the data collection, describes the baseline situation in the area, and goes through the empirical strategies used to estimate the effects of the intervention. Section four describes the results on primary outcomes and provides a number of robustness tests. Section five sheds some light on the potential channels through which the intervention works. Section six makes some back-of-the-envelope calculations of the cost-effectiveness of the intervention before the final section concludes.

2. Intervention and Research Design

2.1 VSLA Intervention

The purpose of the intervention is to encourage the formation of groups with fifteen to twenty-five members who are then trained to manage their own village savings and loan association. As no external capital is provided, the groups are essentially financial intermediaries.

Within the larger microfinance sector, community-managed microfinance belongs to the category of member-based, community-managed, accumulating microfinance institutions (see table 1). Typically, microfinance impact evaluations have focused on the professionally managed microfinance institutions. Unlike these, the VSLAs do not rely on the injection of external funds, just as they do not rely on the sustainability of a professionally managed institution, but rather on the sustainability of the group formed within the local community.

Table 1: Microfinance overview

Ownership	Management	Fund Accumulation	Examples
	Professionally managed microfinance	Large accumulating savings and credit associations (ASCAs)	FECECAM, Benin Savings and Credit Cooperatives (SACCOS)
Member-based institutions	Community-managed microfinance	Rotating savings and credit associations (ROSCAs) Small accumulating savings and credit associations (ASCAs)	Tontines, susu, upatu, merry-go-rounds, chit, pasanakus Village savings and loan associations, Savings for Change, the WORTH model, savings and internal lending communities ¹⁰
For-profit institutions	Professionally managed		FINADEV, Benin Equity Bank, Kenya SKS, India
Non-profit institutions	Professionally managed		Grameen Bank, Bangladesh BRAC, Bangladesh Opportunity International

Source: Own information

¹⁰ Some also include self-help groups, primarily found in India, as savings and loan associations, but they have very different procedures than the ones listed here.

The inspiration for VSLAs came from rotating savings and credit associations (so-called ROSCAs, see Bouman 1995), and was developed by CARE international and VSL Associates during the 1990s (Ashe 2002). The aim has been to improve on ROSCAs in two respects: to make the groups more sustainable and to make them more flexible. Increased sustainability comes from a series of accountability features that prevent theft of funds and elite capture.

One key accountability feature is the constitution, which is developed with the help of a facilitator. It describes areas of potential conflict and their solution—for example lending rules, election procedures, exclusion of members, as well as fines for delays and non-attendance. Another aspect which contributes to transparency in transactions, and thus accountability, is that all transactions take place in the presence of all members at the weekly meetings and are counted independently and in public by two elected money counters. Finally, funds are stored in a cashbox between meetings which is locked with three padlocks. Three different members hold a key to the box, and the box is stored in the house of a fourth member, so no transactions can take place between meetings.

Flexibility is increased because members can at any time borrow the amount they want up to three times their own level of savings—provided that funds are available. Whereas ROSCAs multiply without external facilitation, VSLAs only do so to a small degree—thus requiring facilitation by, for instance, an NGO. This may be due to the complex nature of the accountability features, which will be described in detail below.

In general, as well as in our case, VSLAs are implemented in the following way (see e.g. Allen and Staehle 2007): After conducting awareness meetings in every targeted village, a local NGO facilitates the formation and training of groups. Initially, during the first three months, groups are visited every week to set up procedures. Groups work as member-owned financial intermediaries with three products: savings, credit, and insurance. Savings are compulsory and are collected at the weekly meetings and conceptualized as buying shares. Every week, a member must buy at least one share and is permitted to

buy up to five. The share value is set by the group and written in the group's constitution. It varies between 50 and 100 MK in our case.¹¹

Loans are provided at every fourth meeting. If the funds requested by members exceed the amount of saved funds, the group decides who gets the loan by following a predetermined list of criteria written in the group's constitution. This typically assigns funds based on the stated use of the loan. The interest rate on loans is set by the group, and can thus be used to regulate excess supply or excess demand in the medium run. Usually, the nominal interest rate on loans is set to between five and twenty percent per month, but extensions in repayment schedules and inflation make the real interest rate considerably lower (Rasmussen 2012). Loan contracts run for three months, with a grace period of one month. Rules for loan approval are set in the group's constitution, but most often focus on productive purposes.

The overall interest rate on savings is typically four to five percent per month, but materializes only after the end of a cycle, typically lasting twelve months, when all savings and interest payments are divided by the number of shares and paid out—the so-called share-out (Rasmussen 2012).¹² The interests paid on loans is thus repaid to members as interests earned on savings. At the end of each cycle, members decide whether to leave or remain in the VSLA group, and whether the group should accept new members. Any impact found in our analysis below will thus be the impact of at most two full cycles of collecting savings, giving out loans and returning the savings with interest among the 100 groups established by mid-2011. Usually after one year of initial training and monitoring, groups “graduate,” which means they are no longer supervised by the NGO that helped set them up.

¹¹ Throughout the paper we report monetary values in Malawian Kwacha (MK). 1 USD corresponded to 91.91 MK using poverty-adjusted PPP exchange rates in 2009. See Deaton and Dupriez (2011).

¹² It is lower than the interest on loans primarily due to the fact that not all the funds are lent out all the time and that savings accumulate over time.

Apart from savings and loans, VSLAs also offer insurance that is financed by a small premium paid by each member each week, separately from the savings and credit activities. The insurance is paid out as a grant or an interest-free loan when certain events occur that are outlined in the group's constitution (usually the death of family members, death of cattle, sudden illness, or other emergencies).

2.2 Potential Channels of Impact

In this section we discuss the channels through which the intervention might work. The question is how increased access to the financial services through the intervention might change economic decisions by participants. The services offered by VSLAs are both compulsory and voluntary savings as well as access to loans and insurance. We will focus on each of these channels before briefly commenting on if and how the group meetings in and of themselves could have an effect on participants.

Savings

Savings is the primary and compulsory component of the VSLA intervention. Investigating the channels through which access to new savings opportunities might affect household welfare is therefore central to understanding the overall impacts of the VSLA groups. Members are required to save a small amount every week and can choose to save up to five times this amount. The major cash withdrawal is the share-out, where all funds in the savings groups are distributed. According to the implementation guidelines, members should not withdraw savings during the cycle.

An increase in savings must be driven by an increase in expected utility from savings. This increase in expected utility can be a result of an increase in expected returns if VSLAs offer higher interest rates on savings than alternatives. Another possibility is that VSLAs offer a savings product with lower risk than alternatives, thereby allowing participants to supplement other savings and investment options. Moreover, participants may use VSLAs as a commitment device, thereby avoiding spending money on consumption of unnecessary items. Finally, participants might save in VSLAs simply to get access

to the other features of the group, in particular loans and insurance but also possibly information and informal risk sharing. Below we investigate these options one by one.

In the case where VSLAs offer higher expected returns to savings, short-term behavior will depend on the primary rationale behind saving. However, higher expected returns to savings will spur individuals to save more in order to receive a higher future income, all else being equal. If, as suggested by Deaton (1991), the overarching motive for saving is precautionary, then two things might happen from the intervention: (1) Individuals might save less since the same amount of risk coping can be obtained with a lower savings rate, or (2) Individuals might save more, since risk coping is cheaper and thus they might replace current consumption with future ability to cope with risk. In other words, there is both a substitution and an income effect, and whether the net effect will be positive or negative depends on the relative size of the two. Despite these uncertainties about short-term behavior, all the channels described lead to an increase in income from interests in the long term. This income might be invested, or it might simply be consumed.

Even if the expected returns on savings through VSLAs are not higher, they could still be preferred due to a lower risk than the alternatives. VSLAs can offer lower risk either because the groups are better at monitoring loans than individuals, for example due to improved delegated monitoring, or simply because groups pool the risk on individual loans (Diamond 1984). With concave utility functions, this in itself can improve household welfare. But it can also enable households to undertake more risky investments elsewhere, for example in businesses or agriculture, where expected returns are higher.

Using the VSLAs as a commitment device can be ex-ante welfare enhancing if participants have time-inconsistent preferences, or if the decision maker within the household has other preferences than the participant. One conceptualization of time-inconsistency is the dual self-model, whereby decisions are made through a negotiation between an impatient short-term self and a patient long-term self

(Fudenberg and Levine 2006), but other models exist. Common to all models is the fact that individuals apply different discount rates for the near future compared to the distant future. People with time-inconsistent preferences suffer from self-control problems, and their ability to tackle these problems depends on their beliefs about their future behavior. So-called “sophisticates” foresee their self-control problems and might use commitment devices to overcome them (O'Donoghue and Rabin 1999). Ashraf, Karlan, and Yin (2006) offer a pure commitment device to savers in the Philippines and show that individuals with time-inconsistent preferences are more likely to utilize the product.

However, even if the VSLA participant does not have time-inconsistent preferences, the VSLAs might allow the participant to commit the rest of the household to his or her own preferences. Using the VSLA as a commitment device might enable people to mitigate their spouse's self-control problems (Anderson and Baland 2002)¹³. This might explain why the vast majority of VSLA participants are women: Women decide on day-to-day purchases but might lack the bargaining power to influence larger economic decisions. The commitment component of the VSLAs allow the women to use a fraction of their day-to-day monetary control to increase longer-term savings, assuming they have a say in what to use the share-out money for.

Either way, the commitment component of the VSLAs might enable participants to save larger sums. Since the groups themselves decide on the timing of share-out, they might choose to share out funds at a time where they need lump sums for particular investments (Rutherford 2009). If, for example, share-out is at the beginning of the agricultural season, investment in agriculture might increase. Duflo *et al.* (2011) find that providing farmers with an opportunity to pay for

¹³ Jackson and Yariv (2012) show that, even when an individual's preferences are not time-inconsistent, all households will exhibit time-inconsistent preferences, unless discount rates are equal across household members or decisions are dictatorial (i.e., one household member's preferences decide). As cultural norms would suggest, this would be the husband's preferences, which might be another reason for women joining VSLAs.

fertilizer just after the harvest, when money is available, and having the fertilizer delivered at the time of planting, increases fertilizer use dramatically. In addition, the timing of share-out might serve as a so-called label, fixating the minds of the participants on using the money from the share-out on a particular asset—in this case, fertilizer or other lump-sum investments in agriculture (Thaler 1990). These increased investments in, for example, agriculture can increase productivity and in turn lead to higher household income and an increase in household welfare.

Credit

VSLAs can expand access to credit, for example through more efficient monitoring of contracts or through a more effective pooling of local savings (Diamond 1984, Freixas and Rochet 2008). These efficiency gains can facilitate increased use of credit for investment or other purposes.

An increased access to credit could affect household welfare through a number of channels. The most-used explanation for microcredit to affect household welfare is the story of entrepreneurship: Participants take out a loan for investments that would not otherwise be funded, as illustrated by the model in Banerjee et al. (2010). These investments could be used to expand existing income-generating activities or to start new ones. Increase in access to funds might also lead to specialization, i.e. fewer but larger income-generating activities that exploit economics of scale.

However, expanding access to credit has another possible channel through which it can affect household welfare and which does not depend on households using credit for investment purposes. Since the household is now more confident in its ability to take out a loan if need be, it might switch from ex-ante risk-coping strategies—engaging in less-productive but also less risky activities—to using credit as an ex-post risk-coping strategy. Loans can be used to smooth consumption, for example during the hungry period, if the household is hit by a negative income shock (Zimmerman and Carter 2003). This could cause households to shift their investments into riskier assets

that have a higher expected return as well as a higher risk. Since VSLA members are all from the same area and typically have similar types of income, loans can be used only in the case of idiosyncratic shocks, since the VSLA groups depend on local funds.

An argument against this channel of impact is the fact that VSLAs are instructed to lend out funds for productive purposes only and that this is often written in the constitution of the group. Thus the main impact of the credit channel, if there is one, is likely to be through productive investment.

Insurance

Apart from the indirect insurance through credit and savings, VSLAs also offer a simple explicit insurance product. The groups decide themselves on the exact nature of the insurance product, but it almost always involves insurance against illness and death of household members. As with other insurance products, this might benefit households in two ways, both directly and indirectly: With generally low liquidity, this insurance might be one of the only ways individuals can smooth consumption, thereby achieving an increase in total utility. If there are asset-based poverty traps, where household consumption declines once its assets are below a certain threshold, then households might choose ex-ante coping strategies that involve low risk-low return activities (for example with regard to crops and investment choices, as discussed above). From this perspective, insurance can enable households to choose activities with higher risk and higher expected returns, much like improved access to credit might cause some households to use credit as an ex-post risk-coping strategy (as argued above). This can lead to increased consumption in the long run even though the insurance itself does not pay out (Carter and Zimmerman 2003).

Other Channels

Apart from the savings, credit, and insurance channels already suggested, one can imagine a number of impact channels that do not have to do with the financial products delivered through the VSLA

groups. During the awareness campaign, it was stressed that participants should form groups with people they trust, but participation might in itself also build trust. This trust might, in turn, facilitate lower costs of informal contracting in joint business activities, informal lending, or informal insurance arrangements against idiosyncratic risks. Another channel of impact could be increase in information flow in the groups, which likewise could lead to a higher degree of informal insurance, the spread of new and more efficient agricultural techniques, or other types of interaction. As it is often the case that women are more likely to participate in the VSLAs, this can also result in a general empowerment of the women in the area. An increased female empowerment might affect village decisions, as it has been found that gender affects policy decisions (Chattopadhyay and Duflo 2004). Finally, implementers report that groups often discuss issues of relevance to their villages more broadly. This might lead to spill-over effects, as treatment villages might be more effective at solving collective action problems.

2.3 Design of the Experiment

The crucial challenge for an impact evaluation is to construct a credible counterfactual that is not sensitive to selection bias (Duflo *et al.* 2007, Angrist and Pischke 2009, Banerjee and Duflo 2009). Two types of selection bias are particularly relevant in our case: non-random program placement and self-selection into program participation. The roll-out of programs like ours is generally far from random. Indeed, from conversations with implementers it was clear that some villages were more eager than others to start the program. Village chiefs or eager villagers lobby field officers to come to their village and train VSLAs. Moreover, it seems likely that the individuals who self-select into any microfinance program are different from the general population in important ways (see the following chapter of this thesis).

We address the problem of non-random program placement by randomizing the roll-out of the VSLA intervention at the village level. As will be described in further details in the empirical strategy below,

this allows us to estimate the intention-to-treat effects (ITT) as well as local average treatment effects (LATE). The ITT is intuitively the effect of introducing the intervention to the households from the villages that were allocated to the treatment group, whereas and the LATE is the effect on the households from the treatment villages who were actually induced into participating in the intervention due to the randomization. While it might seem more fruitful to randomize at the individual level to estimate the more general average treatment effects¹⁴ and average treatment effects on the treated,¹⁵ this is impossible given the nature of the intervention. The intervention relies on villagers forming the groups themselves to ensure a high level of trust among group members from the start. Randomizing at the individual level would interfere with this crucial part of the implementation.

Out of forty-six villages in the program area, we randomly chose twenty-three villages for implementation in the first year (the treatment villages) and twenty-three villages for implementation in the third year (control villages). As such, the study is a simple parallel trial—that is, only one treatment group and one control group. In randomized control trials, balance in observable and unobservable characteristics can be improved in a number of ways. This is particularly important in our case, given the relatively low number of units upon which the randomization is carried out (Bruhn and McKenzie 2009). One way of improving balance is to randomize, test for balance, and rerandomize until some pre-specified level of balance is achieved. This method has been criticized, as it is not clear how to take analyze the data after the rerandomization (Bruhn and McKenzie 2009).

¹⁴ The average treatment effect (ATE) is the expected effect of treatment on a randomly chosen individual from the population. This is often used in observational studies and not in randomized control trials, where we do not know what the effect of treatment would be for the part of the population that does not participate, despite being allocated to the treatment group.

¹⁵ The average treatment effect on the treated (ATT) is the expected effect of treatment on a randomly chosen individual from the population that has undergone treatment. This is closely related to the LATE effects, which we comment on in the empirical strategy section.

Another strategy is to stratify the random sampling, where units are grouped into strata, or blocks, based on some characteristics believed to be correlated with central outcomes of interest (Bruhn and McKenzie 2009).¹⁶ By ensuring that an equal number of villages within each of these blocks belong to the treatment and control group respectively, the balance in baseline characteristics of the households across the randomization can be improved. Bruhn and McKenzie show that in smaller samples, and under certain conditions, blocking (and also pair-wise matching) out-performs the commonly used praxis of re-randomization.

We chose stratification based on the above considerations—also, in part, since this allowed closer collaboration with the implementing partner. By allowing the field officers who implemented the project to execute the randomization under our supervision, we enhanced their understanding for the research strategy and the importance of complying with the randomization. Prior to the randomization, we identified general characteristics of the villages and allocated them to seven non-overlapping blocks. The blocks were defined as follows: large fishing villages, small fishing villages, particularly eager villages,¹⁷ large non-fishing villages, villages with a rice irrigation scheme, villages with another NGO-led intervention, and a final group of the remaining villages. Finally the randomization was carried out, by the field officers drawing village names from seven hats each containing the villages of one subgroup under our supervision. Figure 1 below shows the physical location of the village centers, with the shape of the dot indicating the block to which the village belongs, and the shape surrounding each village center indicating allocation to the treatment or control group.

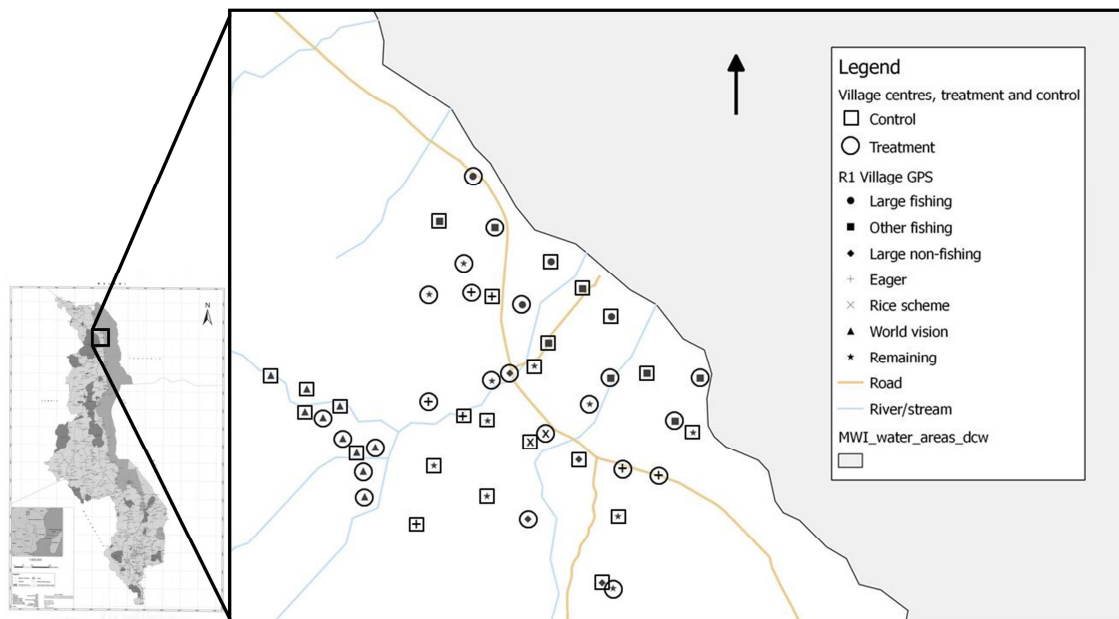
As part of the VSLA intervention, an awareness campaign was carried out by the implementing partner in all forty-six villages prior to randomization. We used these awareness meetings by having the

¹⁶ We use the word “blocks” here to differentiate the stratification of villages from the stratification of sampled households.

¹⁷ These were identified by the field officers based on the reaction from villagers at the awareness meetings.

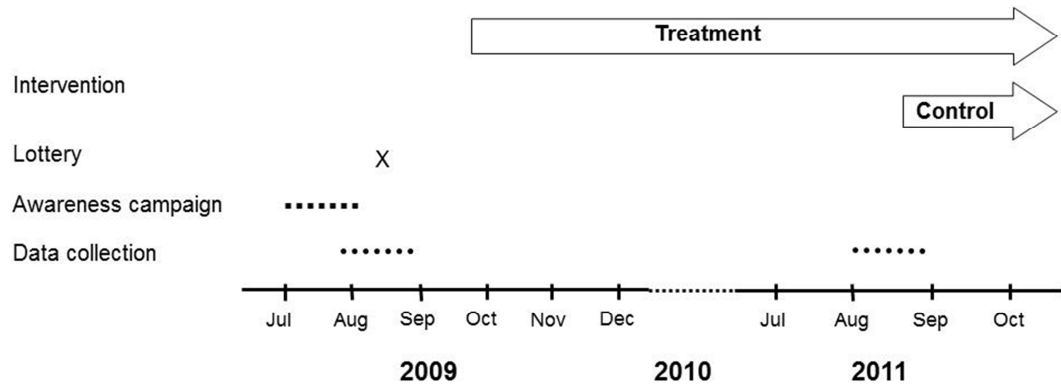
field officers from the implementing partner collect lists with the names of the villagers who expressed an interest in joining the VSLA groups. These lists of interested households were potentially beneficial for the research, as will be described in the section below on the empirical strategy.

Figure 1. Villages and Randomization Blocks



The following figure provides an overview of the timing of the randomization, implementation, and the data collection which—along with the sampling strategy and its implications for the analysis in the paper—will be described in the data section below.

Figure 1: Timeline



Minimum Detectable Effect

The aim of any impact evaluation is to provide a credible answer as to whether an intervention had an effect or not. Basing this assessment on quantitative methods creates the possibility that an impact evaluation provides a false negative or Type II error; that is, finding no change when in fact there is one. This can happen if the data collected is not suitable for detecting the changes in the outcomes, in particular if the sample size, and thereby the statistical power of the experiment, is too small. Since the randomization credibly overcomes problems due to non-random program placement, it is important to minimize the risk of inadequate data or inefficient design of the experiment driving any lack of significant results.

This has made power calculations an important part of the randomization toolkit (Duflo *et al.* 2007). Using characteristics of the randomization, while making assumptions on the variance of outcomes of interest as well as the expected size of a program's impact, the power calculations estimate the sample size required to detect statistically significant changes in outcome variables. More importantly, the power calculations allow the researchers to compare different research designs and identify the strategy that provides the greatest power for detecting effects of the intervention. In this section, we provide power calculations based on baseline data to suggest the

required size impacts must have to be detectable in the current research design.

In a case of simple randomization at the individual level, the power calculations determine the minimum detectable effect (MDE). This is the effect size that, if it is true, will have a κ probability of leading to a positive significant result in an impact assessment at α significance level. κ is called the statistical power of the assessment, and usually ranges from 80% to 90%, while α is the significance level, for example 5% (Bloom 1995). Bloom argues for using one-sided t-tests, since there is an a priori understanding about the direction of any effects from the intervention, although this is not standard in economic research. Based on these significance levels and with assumptions on the variance of the outcome of interest, σ ,² the MDE for individual randomization can be calculated using the following formula:

$$MDE^{IR} = (t_{1-\kappa} + t_{\alpha}) \sqrt{\frac{\sigma^2}{P(1-P)N}}$$

where P is the proportion of units allocated to treatment, $t_{1-\kappa}$ and t_{α} are t-values using the design degrees of freedom and N is the sample size. However, our research design differs from this basic setup in a number of important dimensions. First of all, our unit of randomization is the village level, whereas the unit of observation is the household. Second, we randomize within blocks. And third, we have panel data available. While the latter two are not easily incorporated into power calculations, randomization at the group level is rather standard, and tools for calculating the MDE in cluster-randomized trials are well-developed.

Following Duflo et al. (2007), we take the randomization at the group level into consideration by adding a design effect to the above power calculations. The MDE for group randomization is calculated as:

$$MDE^{GR} = D \cdot (t_{1-\kappa} + t_{\alpha}) \sqrt{\frac{\sigma^2}{P(1-P)N}}$$

where $D = \sqrt{1 + (n - 1)\rho}$ is the design effect, and $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ is the intra-cluster correlation where τ^2 and σ^2 are the between- and within-group variations respectively.

Based on information from the baseline survey, we estimate the intra-cluster correlation, ρ , as well as the standard deviation for each of the predefined outcomes, to estimate the MDE given our research design and survey size. Even though Bloom (1995) argues for using one-sided t-tests, we estimate the MDE for both one-sided and two-sided tests in table 2 below. The variables are explained in more detail in the next section.

Table 2. Minimum Detectable Effects (Power Calculations)

Outcome	N	Mean	Std. Dev.	ρ	Design effect	SE(beta)	MDE	
							One- sided	Two- sided
Number of months with fewer than three meals a day	1732	4.10	4.03	0.11	2.22	0.19	1.08	1.23
Number of meals yesterday	1732	2.65	0.56	0.06	1.79	0.03	0.12	0.14
Food consumption per adult equivalent per day (MK)	1732	75.62	48.69	0.03	1.51	2.33	8.91	10.10
Number of income-generating activities (including agriculture and livestock)	1732	1.99	1.10	0.03	1.40	0.05	0.19	0.21
Total savings	825	7,385	31,755	0.02	1.31	1,519	5,029	5,699
Savings in VSLA	825	186	2,360	0.02	1.28	113	368	417
USAID PAT consumption per capita per day (MK)	1732	69.95	32.50	0.04	1.58	1.55	6.22	7.05
Size of house (number of rooms)	1732	2.75	1.25	0.05	1.66	0.06	0.25	0.29
House has cement floor	1732	0.10	0.30	0.04	1.56	0.01	0.06	0.06

The final two columns show the minimum detectable effect based on one- or two-sided t-tests. This should be compared to the baseline mean, shown in the second column, in order to assess the size of the effect required for us to detect it. Finally, it should be noted that the above calculations do not take the randomizations within blocks into account, just as we expect to decrease the MDE since we have panel information, which allows us to decrease the variance on the estimated effects. (See discussion below.)

2.4 Planned Objective and Predefined Outcomes

Above we presented the possible channels through which the VSLA might work. We did not, however, specify the final outcomes and how these should be measured. That is the goal of this section. In particular, we will describe how the outcomes were predefined through the use of a so-called logical framework matrix. We will also discuss the gains from using predefined outcomes.

Using outcomes defined prior to the analysis limits the chance of data mining. Data mining occurs when researchers, perhaps aware of publication biases (DeLong and Lang 1992), search for significant results among a wide range of possible outcomes. Conscious or unconscious data mining can be a serious problem: If a researcher tests ten null hypotheses for ten independent variables, each with a 5% chance of a Type I error (finding a relationship when there is none), then the overall probability of a Type I error is 40% (Duflo *et al.* 2007, Rasmussen *et al.* 2011). Adjustment of p-values is one solution to this problem, although this requires researchers to credibly reveal the total number of outcomes tested, as it would be insufficient to adjust for the number of outcomes reported. Pre-specification is another, more transparent solution: If the outcome variables and their measurement are predefined, then the initial p-values are still valid. However, since we carry out multiple tests, adjustment of the p-values would still be justified. We do not adjust for two reasons: First, there are a number of arguments against adjustment, so it is not currently standard practice in economics (Gelman *et al.* 2012). Second, we would have to arbitrarily partition our outcomes into “families of

outcomes.” Nevertheless, the fact that we report results on all predefined outcomes enables readers to take the issue of multiple outcomes into account.

Presently, J-PAL maintains a public “hypothesis register” with four registered trials, and the American Economic Association is planning a trial register. However, at the time when this analysis was designed, there was no trial register in place that would accept trial registration of non-medical trials.¹⁸ Instead, we used the so-called logical framework matrix to pre-define our outcomes. We also report extensively on non-predefined outcomes in order to investigate the channels of impact. By doing so, we explicitly distinguish between primary and secondary analysis.

The logical framework matrix is a part of the logical framework analysis, by far the most common project-management tool in international development (Dale 2003). The matrix is made by the implementing partner prior to initiating a project and specifies the expected impacts of the intervention. Since the matrix is made by practitioners rather than researchers, it is not very detailed when it comes to measurement and analysis. For example, even though the matrix mentions indicators on which the intervention is expected to have an impact, it does not specify how each of these indicators is to be measured. Therefore, it does not serve the purpose of a pre-analysis plan used by, for example, Casey *et al.* (2012). Furthermore, the matrix is not well suited for our cluster randomization framework, in that it lists goals specifically for participating households. Our research design does not allow us to uncover the effects for participating households, as will become evident in section 3.4 below. Instead, we analyze the intention-to-treat effects among all households in the treatment villages. We do not look at how many households

¹⁸ ClinicalTrials.gov and related registers accepts only trials with medical outcomes. Other registers available are not considered trial registries in the sense referred to above, as they do not enable users to see when initial hypotheses were registered. This includes the databases by OECD DAC and 3ie as well as the What Works Clearinghouse registry (education) and the C2-SPECTR registry (criminology).

experience an increase, but simply if there is an increase on the average among all households in the VSLA villages.

The first two columns of table 3 below show the expected outcomes and indicators listed in the logical framework matrix by the implementing partner, Soldev, prior to the project being initiated. Based on this information, we developed sections of the questionnaire that could quantify each of the indicators. The third column lists the questionnaire variables used for the primary analysis, numbered one to ten.

As is clear from the table, there is some freedom as to how to interpret most of the outcomes in the logical framework matrix, and in two cases, two variables are chosen to capture one outcome. These choices were made based on validity and measurability. This, and the fact that at the time, there was no publically available registry, is the reason why our analysis is not prespecified.

Table 3. Predefined Outcome Measures

Logical Framework Outcome (Soldev)	Logical Framework Indicator (Soldev)	Corresponding Questionnaire Variables (Researchers)
Increased food security	Hunger period is reduced	Hungry period measured as the number of months with fewer than three meals per day.
	Increase in the consumption of food	Number of meals yesterday. Food consumption measured through recall of 17 food items.
Increase in income-generating activities	The average number of IGAs carried out by the VSLA participants has increased	Number of enterprises run by the household.
	Increase in the volume of savings by the VSLA groups from project related activities by 2012	All highly liquid savings Savings in VSLAs
Improved household income	The share of the targeted population living below 1.25 USD/day has decreased as measured using USAID's PAT	USAID PAT's prediction of per capita consumption
	HHs have improved their housing standards	Number of rooms in dwelling House has cement floor
	Increase in household asset ownership	Land ownership (acres)

All variables are measured using the survey described below. Some of them require explanation. The hungry period is measure through recall, where we ask about average meals for the household in each month over the last year. Food consumption is measured using weekly recall of seventeen food items as described in Appendix 1. These items were also included in the 2004/05 Malawi Integrated Household Survey and are the most important food items according to this survey. They make up eighty-nine of total food consumption.

All highly liquid savings include savings at home, with friends, in banks, in microfinance institutions and in savings groups. USAID PAT is a predictor of total consumption based also on the 2005/5 Malawi Second Integrated Household Survey. It is different from food

consumption in two ways: First, it is an indirect measure with twenty questions selected on the basis of their ability to predict total consumption as well as being easy to measure (IRIS Center 2012). Second, it measures total consumption, not just food. Total consumption is the value of all goods and services included in the 2004/05 Malawi Integrated Household Survey, including housing, health, transport and clothes (World Bank 2004). In this situation, consumption is used as a welfare measure (Deaton and Grosh 1998), and for that reason it is listed in the income section.

3. Data and Estimation Strategy

3.1 Data

Data was collected with the primary purpose of conducting the analysis in this paper. However, the opportunity to collect panel data from rural Malawi spurred a number of other research questions that could be assessed. Hence, the questionnaire was designed to accommodate these research questions as well. The following describes the design of the questionnaires, the sampling strategy and its implications for the analysis carried out, and the actual data collection, including the tracking undertaken to limit the attrition.

Questionnaire Design

Four different questionnaires were developed for the data collection: a village questionnaire, a household-head questionnaire, a spouse questionnaire, and a short questionnaire. The latter three were all administered to household members to obtain either personal or household-specific information. The inclusion of separate questionnaires for the household head and the household head's spouse allowed for unique data on intra-household matters and person-specific information.¹⁹

The questionnaires were developed in English and subsequently translated into the local language, Tumbuka, by the Invest in

¹⁹ We experienced no cases where a woman was a head of the household, except for instances where women were living without a spouse.

Knowledge Initiative (IKI), that was also responsible for the data collection. The translations were checked by the group of interviewers as part of their training prior to the data collection.

To answer a wide range of research questions, two different questionnaires were developed for each household: one intended for the head of the household and one for the household head's spouse. The reason for having two questionnaires per household was twofold: First, it allowed for questions related to intra-household issues, and second, it divided the time spent on answering household-specific questions between the head and the spouse. For the latter reason, many sections appeared either in the head questionnaire or the spouse questionnaire, but not in both. The division was made according to who would most likely have the necessary information and was finally decided during piloting of the questionnaires. For instance, the piloting made clear that accurate information on the age of the household members was most easily collected from the spouse, while information on assets and general household characteristics was more accurately answered by the household head. In situations where the household head was an unmarried female, she answered both the household head and spouse questionnaires, although, to minimize interviewer fatigue, these two interviews were done separately.

The short questionnaire was an abbreviated version of the head and spouse questionnaires, and included primarily the questions necessary for assessing the impact of the intervention on the predefined outcomes. The short questionnaire was developed to maximize the power of detecting any impacts of the project, given the budgetary restrictions on the data collection.²⁰ Households were randomly allocated to answering the set of household head and spouse questionnaires or the short questionnaire following the sampling strategy described below.

The village questionnaire was administered to a group of knowledgeable individuals from each village to obtain village-specific information on topics like schools, markets, health facilities, banks,

²⁰ See the implication of sample size in the above discussion of minimum detectable effects.

and so forth. It also included information on the GPS—location of the village, number of households in the village, typical economic activity of the villagers (such as the primary types of crops grown), any recent exposure to shocks, as well as information on prices.

All questionnaires were piloted and adjusted accordingly during two weeks of field work prior to the first data collection. The final round of data collection was preceded by a shorter period of piloting, focusing on the new modules that were included in this later survey.²¹ The piloting was conducted by the authors together with interviewers from IKI in villages surrounding the actual survey area. This ensured a similar environment as could be expected in the actual survey area, without interfering with the respondents within the actual survey area. Importantly, the main local language in the pilot area was Tumbuka, allowing for any problems in understanding and interpretation of the questions to be detected and fixed before actual data collection took place.

Sampling

Since the primary object of interest was assessing the impact of the introduced VSLAs, steps were taken to maximize the power of the experiment with regard to this goal. We stratified the sampling within each village by whether the household had declared an interest in participation, using the information gathered by Soldev during the awareness meetings as described earlier.²² This was done to enable oversampling of households that had expressed interest in participating in the upcoming groups.²³ We sampled roughly the same number of households in all villages. Due to differences in village size, this led to considerable variation in sampling probability between strata apart from the oversampling of interested households. Table 4

²¹ These modules gathered detailed information on the use of the VSLA groups as well as the monetary spending patterns of the household head and his spouse.

²² The actual sampling of households was done by generating a random number between 0 and 1 for each household using Stata, and assigning the X households with largest numbers from each strata to be interviewed.

²³ In the section on the empirical strategies we explain how an analysis on these interested households might have allowed us to estimate the ATT.

shows the mean sampling rate of interested and non-interested households as well as the standard deviation of the sampling rate within these strata across villages. Had we sampled proportionally within villages, the standard deviation would have been zero.

Table 4. Stratum Sampling Rates

	Number of Strata	Mean	Standard Deviation (percentage points)
Interested	46	68%	14
Non-Interested	44	24%	17

Note: The table shows that the stratum sample rate is not the same across strata.

With no prior knowledge about the variation within strata, it is likely that the optimal sampling rate would have been proportional allocation since our primary interest is to compare the average outcomes of villages allocated to treatment and control. Thus, through our interviewing an absolute rather than a proportional number of households from each stratum within the villages, the variation in sampling rates might increase the standard errors of the estimates (Parsons 2005).

Weights

In order for the results to be informative of the entire area, we used sample weights to correct for the stratified random sampling of households. Specifically, households were weighted according to their inverse probability of being sampled to allow us to generate consistent estimates of the population regression function (Angrist and Pischke 2009). The survey literature seems to agree with this approach. Pfefferman (1993:323) concludes that “failure to account for all the important design variables or incorrectly specifying the conditional distribution of the survey variables given the design information can have severe effects on the inference process. These effects include bias of point estimators and poor performance of test statistics and confidence intervals.” With our design, we have information about the

data generation process that we use in estimating population parameters.²⁴

We used sampling weights in all the analysis shown below. However, as a robustness test, we also display the un-weighted regression results for the predefined outcomes.

Data Collection

Data was collected in 2009 and 2011 by the Invest in Knowledge Initiative under the supervision of the authors. In 2009, data collection took place from July 26 to August 30. The endline data was collected between July 8 and August 14, 2011. Given that the primary source of income is agriculture, one of the concerns when conducting a survey in rural Africa is ensuring compatibility with the agricultural seasons. The harvest in Northern Malawi takes place from April (the “green harvest”) to July, but was in both years completed in the villages prior to data collection. This ensured availability of households to participate in the interviews as well as the possibility of gathering information on the harvest based on relatively short recall periods.

In both the 2009 and 2011 data collection, 24 interviewers completed the approximately 2,600 interviews—834 interviews with the household head and spouses respectively and 950 short questionnaire interviews. In the 2009 data collection, the interviewers were divided into three teams with a supervisor assigned to each of these teams. During the 2011 survey, the interviewers were divided into four teams to more easily accommodate the tracking of migrated households as will be described further below.

²⁴ Despite the agreement in the survey literature and part of the econometrics literature, the issue of weights warrants further discussions. Weights affect results greatly, and the econometric literature outside of survey research does not agree on the issue, at least at a first glance (Angrist and von Pischke, particularly pp. 91-94). If the population regression function is interpreted as causal in all its parameters, as would be the case in structural equation modelling, then weighting is in some cases not justified regardless of the underlying method of sampling (Cameron and Trivedi 2005). If, however, there is only one parameter of interest, for example because the causal interpretation stems from exogenous variation rather than from a correct structural specification, weights should be applied. This is the case within a counterfactual framework, or more specifically the Rubin causal model (Rubin, 1974, Wooldridge, 2001, Angrist and von Pischke, 2009).

Interviewers and supervisors were trained in the specific questionnaires for approximately one week prior to actual data collection. This training was done by the researchers in the traditional authority (TA) bordering the survey TA. Training included a detailed walk-through of the questionnaires, mock interviews of individual sections by interviewers on interviewers, and actual test interviews with respondents from the surrounding villages, ensuring no village from the actual survey area was visited during this training.

The head and spouse questionnaires took approximately two hours each to complete, and the short questionnaire approximately forty-five minutes. As a compensation for the time spent by the respondents completing the interviews, bars of soap were provided after completion of the 2011 survey, but not after the 2009 survey. The respondents were not aware of the compensation prior to the first interview, hence this could not serve as a possible factor in any non-response.

The following figure shows the geographical location of the surveyed households. The colors of the dots indicate whether the household belongs to a treatment village or a control village; a circle indicates a household from a control village, and a cross is a household from a treatment village. The resulting picture is quite messy with crosses and circles intertwined. The reason is the spread-out nature of the villages in this rural part of the country. This has implications for the compliance with the outcome of the randomization, as it makes it possible for some households from control villages to participate in VSLA groups in the treatment villages. The sections on take-up and the empirical strategies below elaborate on the implications of this for the analysis in the paper.

Figure 2: Location of Surveyed Households



Non-Response, Refusal, and Replacement

Some households sampled could not be interviewed. The primary reason was the inability of interviewers to locate the household due to the village leaders not knowing the designated respondent. Given the close-knit nature of the rural communities, it is probable that this situation arose due to problems with the census lists provided by the local agricultural extension office located in Karonga. In some cases the census lists were simply outdated, with the villager sampled for an interview being deceased, having moved away, or simply not being recognized by the village leaders. Finally, a few respondents refused participation in the survey.

In all these cases, the sampled households were replaced using a list of replacement households sampled by the authors. These

replacement lists were made separately for the interested and non-interested households following the initial sampling strategy described above.

Attrition and Tracking

To limit the attrition, the 2011 data collection included a number of initiatives to survey as many of the households as were part of the initial sample. Furthermore, care was taken to ensure that the 2011 survey respondents would be the same as the respondents from the 2009 survey—that is, no other members from the household answered the questionnaires unless necessary. Supervisors visited the villages prior to actual interviewers entering the village to make appointments with as many of the respondents as possible. If respondents were not found at home despite appointments, interviewers and supervisors were instructed to re-visit the household a total number of three times. Only at the third visit where the designated respondent was still not present could another individual from the household be interviewed.

Households that had moved—or households where parts of the group had moved out (such as on account of divorce)—were tracked in the following way. A tracking section was created for the 2011 questionnaires, which recorded information about the new location of the household/household member as well as any contact information, such as phone numbers. If the entire household had moved, this information was gathered from neighbors or village leaders—whoever had the most information about the new dwelling of the household. Any household or household member that had moved within approximately one hour's drive from the survey area was tracked by one of the four teams of interviewers, which were transformed into a special tracking team approximately halfway into the 2011 data collection. Households that had moved outside this area—or where there was no information from the old neighbors about the household's current whereabouts were not contacted. For example, the untrackable household may have been chased out of the village for fear of witchdoctors.

As a result of these initiatives, a total number of 1,775 households were surveyed in 2011. In other words, the attrition between 2009 and 2011 was only 49 households or less than three percent of the initial sample. Compared to other panel surveys, this is a low number (Glewwe and Jacoby 2000). When the number in the regression is lower than 1,775, this is due to non-response on a particular variable.

Since we also tracked households where one of the designated respondents had moved out (for example due to divorce), we have a number of split households where one household from the 2009 survey had become two households in 2011. In the estimations below, we dropped one of these new households randomly to ensure a balanced panel. We assess the implications of this for the results in the robustness section by excluding all split households.

3.2 Baseline Balance

The randomization of villages into our treatment and control groups should, in principle, result in two identical groups. However, given the relatively low number of villages on which the randomization took place, we might simply by chance have created groups of villages that are not comparable. The current section therefore investigates the characteristics of the surveyed villages at the time of the baseline data collection in 2009—that is, before the implementation of the intervention.

Table 5 below provides descriptive statistics on the characteristics of the households as well as our predefined outcome variables and some additional intermediate outcomes, which will be described in further detail below, for the entire sample of households in 2009. The average household was headed by a thirty-nine-year-old man with just under seven years of education. It had six household members, who at the time of the interview consumed approximately 1.17 2005 USD per person per day using the USAID PAT measurement ($e^{0.16}$), and had more than four months of the year where they consumed fewer than three meals a day. Throughout the analysis, we used the 2005 poverty-adjusted purchase power parity exchange rates suggested by Deaton and Dupriez (2011).

Columns four and five report the mean value of the variables for the households in the treatment villages and control villages respectively. Column six reports the standard errors from a test for any significant differences between the two means identified by running an OLS regression similar to the one we use later in the impact analysis. Apart from a dummy for treatment village, it includes a vector of randomization blocks as well as sampling weights. On our predefined outcomes, only one variable differs between the treatment and control villages: land ownership. Furthermore, food consumption is imbalanced when we do not use logged values. This is likely to be due to outliers.

Given the randomization of forty-six villages, we would expect some significant differences to occur between households in treatment- versus control villages. We conclude that the baseline is balanced.

Table 5. Baseline Characteristics and Balance

Variable	N	Mean	SD	Treatment Average	Control Average	Difference (t-value)
<i>Project Outcomes</i>						
Number of months with fewer than three meals a day	1732	4.10	4.03	4.26	3.92	1.20
Number of meals yesterday	1732	2.65	0.56	2.61	2.70	-1.49
Total food consumption per week per adult equivalent no infl (MK, log)	1732	6.10	0.58	6.06	6.14	-1.56
Number of income-generating activities (including agriculture and livestock)	1732	1.99	1.10	1.94	2.04	-1.66
Total savings (log)	825	5.70	3.80	5.98	5.40	0.73
VSLA savings (log)	825	0.17	1.16	0.21	0.13	1.13
USAID PAT per capita consumption (log)	1732	0.16	0.42	0.15	0.16	-0.40
Size of house (number of rooms)	1732	2.75	1.25	2.73	2.77	-0.34
House has cement floor	1732	0.10	0.30	0.11	0.09	1.48
Land ownership (acres)	1732	2.70	2.35	2.49	2.92	-3.01***
<i>Other Household Characteristics</i>						
Food consumption per adult equivalent per day (MK)	1732	75.64	48.62	71.54	80.06	-2.24**
Age of household head	1721	38.98	15.33	39.06	38.89	0.24
Number of household members at time of interview	1722	5.77	2.46	5.72	5.83	-0.61
Household is female-headed	1732	0.16	0.36	0.17	0.14	0.89
Years of education of household head	1729	6.86	3.25	7.06	6.65	1.56
Household owns land	1732	0.96	0.19	0.95	0.97	-1.48
Household is member of VSLA	1729	0.06	0.23	0.06	0.05	0.58

(Table continues on the next page)

Variable	N	Mean	SD	Treatment Average	Control Average	Difference (t-value)
<i>Agriculture</i>						
Household has any irrigated plots	823	0.28	0.45	0.22	0.34	-1.79*
Received any coupon(s) for fertilizer	824	0.57	0.50	0.54	0.60	-1.02
Received any coupon(s) for seeds	824	0.53	0.50	0.55	0.51	0.56
Household uses any fertilizer	817	0.70	0.46	0.70	0.70	-0.39
Household uses any fertilizer on maize	1723	0.45	0.50	0.48	0.42	0.76
Household used any purchased seeds	817	0.80	0.40	0.78	0.82	-1.10
Total value (MK) of harvest (all crops)	817	65215	9195	64132	66367	0.20
Total value of maize harvest (MK) using median prices in area	1723	11522	1326	11322	11736	-0.44

Note: Standard errors are clustered at the village level and based on weighted regressions. Displayed results are on the sample of 1,732 households for which we have information for all prespecified outcomes; i.e. the sample used for the pooled difference-in-differences estimations below, unless values are missing on the specific variable. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.3 Take-Up

The following section documents the take-up of the intervention in the survey area. The intervention was effective in starting up VSLAs and attracting households to participate. However, not all households living in villages assigned to treatment participated, just as some households from the control villages found their way into a VSLA group. In other words, there is two-sided non-compliance (Gerber and Green 2012). This has some implications for the estimation strategy and the type of treatment effects that can be estimated. We will briefly comment on this in the current section, but expand on the implications in the section on empirical strategy.

Membership increased during the 2009–2011 period, and the levels were substantially higher at the 2011 survey in the treatment villages

compared to control villages. The situation is summed up in table 6. At the time of the baseline survey in 2009, five percent of the population in the control group and six percent in the treatment group reported that they were members of a VSLA or a similar savings group. Two years later, the figure was fifteen percent for the control group and thirty-nine percent for the treatment group. This increase in participation over time is significant in both treatment villages and control villages, and the difference between the take-up of the VSLA intervention between the treatment versus control villages is significant at the 1% level. In other words: The randomization was effective in inducing villagers to participate in the VSLAs.

Table 6. VSLA Membership

	Control Villages	Treatment Villages	Differences
Baseline (2009)	5.5%	6.2%	0.7%
Endline (2011)	15.1%	39.1%	24%***
Differences	9.6%***	32.9%***	23.3%***

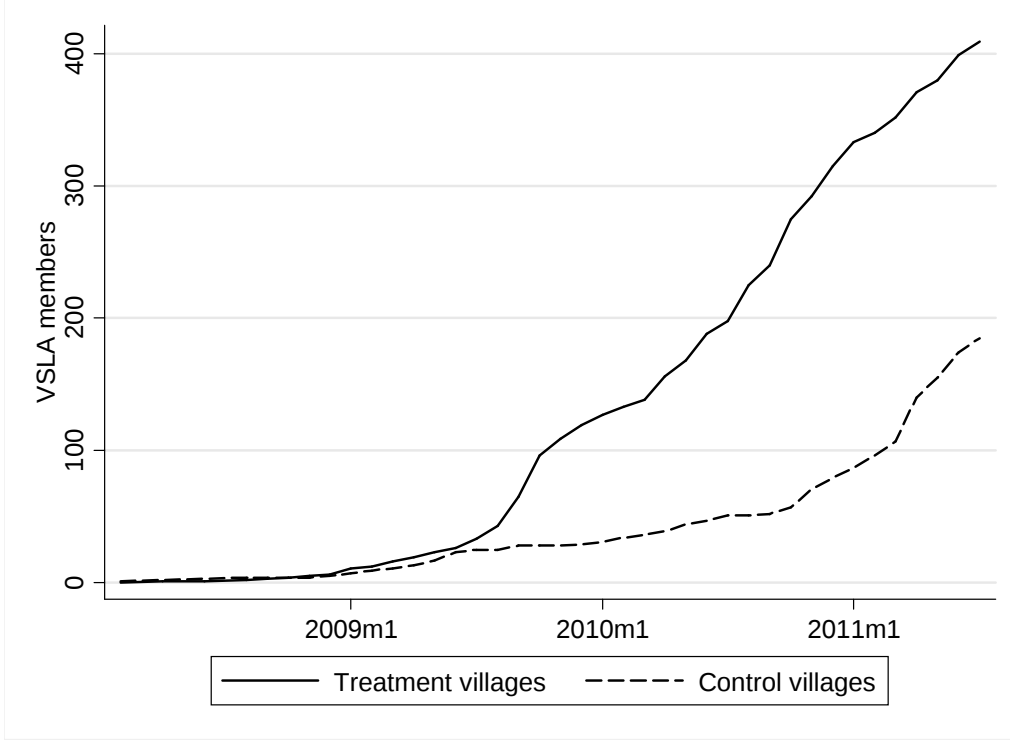
Note: Standard errors used in calculating significance are clustered at the village level. *** indicates significance at the 1% level.

While the randomization induced households from treatment villages to participate, compliance with treatment was far from perfect. VSLA membership in control villages, i.e. contamination or always-takers, can come from the fact that village borders are adjacent and distances between village centers are typically less than two kilometers. Furthermore, as was evident in Figure 2 showing the location of surveyed households, villages are quite dispersed, hence a household might not be located much farther from the village center of a neighboring village than from the center of its own. People from one village might join VSLAs in the neighboring village. Another explanation for the contamination could be that groups in control villages started spontaneously or with the aid of VSLA members from treatment villages.

On the other hand, there is also non-compliance in the treatment villages: more than half the households in the treatment villages were not members of a VSLA at the time of the endline survey. This two-sided non-compliance will lower the statistical power of detecting an impact at the village level and thus increase the minimal detectable effect at the household level, as noted by Duflo *et al.* (2007). In the empirical strategy below, we provide a couple of suggestions for taking this non-compliance into account when we estimate the treatment effects.

In terms of effects, the time of membership matters. As discussed above, some of the channels through which VSLAs might affect households are slow, others faster. In the survey, we asked people if and at what time they joined a VSLA, and Figure 3 shows cumulative membership of VSLAs in treatment villages and control villages during the period. Until mid-2009, there was a slow but general increase of membership in the area regardless of the random assignment into treatment and control. This is probably due to the general popularity of VSLAs, some of which were promoted on a small scale by other NGOs in the same area. In 2009, when the intervention commenced, membership took off in treatment villages. In control villages, however, membership seemed to follow the general pre-project trend until late 2010. In summary, this seems to indicate that control group contamination happened relatively late and thus that the effect of the project on the control group is likely to be small.

Figure 3: Cumulative VSLA-Membership



3.4 Empirical Strategies: Intent to Treat

The randomization allows for a simple estimation strategy in order to estimate the treatment effects of introducing VSLAs to the villages. Following Duflo *et al.* (2007), let $Z = \{0,1\}$ indicate whether village j was assigned to treatment and let $Y_{ij}^{Z=z}$ be the outcome of household i in village j , given its treatment status z . Using the language of expected outcomes, the randomization ensures that $E[Y_{ij}^{Z=1}|Z=0] = E[Y_{ij}^{Z=0}|Z=0]$ and thus the treatment effect can be calculated as $E[Y_{ij}^{Z=1}|Z=1] - E[Y_{ij}^{Z=1}|Z=0] = E[Y_{ij}^{Z=1}|Z=1] - E[Y_{ij}^{Z=0}|Z=0]$. In other words, by randomizing which villages receive the intervention, we ensure that the expected outcomes in treatment villages and control villages are similar in the absence of treatment. Specifically, this ensures that there is no selection bias due to the placement of the program: The expected outcomes for individuals in the treatment villages, $Y_{ij}^{Z=1}$, would have been the same as the expected

outcomes in the control villages, $Y_j^{Z=0}$, had they not been exposed to treatment.

Since we do not have perfect compliance, we can estimate the intention-to-treat (ITT) effect (that is, the average effect of introducing the VSLA intervention on all the households in the villages where the project was introduced, irrespective of whether they actually participated).²⁵ A consequence of this is that we can estimate the intention-to-treat effect as $\delta = \hat{E}[Y_{ij}^{Z=1}] - \hat{E}[Y_{ij}^{Z=0}]$ where $\hat{E}[\cdot]$ is the sample analogue of $E[\cdot]$.

In designing the research strategy, we foresaw this potential non-compliance with the randomization. This caused us to collect information from the awareness meetings held by SOLDEV about which households were interested in joining the upcoming VSLA groups, as described earlier. In effect, the households present at the awareness meetings were asked to sign up on lists provided if they wanted to join the intervention. If these lists of interested households could be used to identify the households that later joined the VSLA, we could use the lists to compare the interested households from treatment villages, that actually joined the VSLA, with the interested households from the control villages, that were unable to join a VSLA. If all interested households in treatment villages had eventually joined a VSLA, this would have identified the average treatment effect on the treated (provided the control households complied with the randomization and did not join groups in other villages). However, only about half of the households from the treatment villages listed as being interested ended up being a member of a VSLA group two years later. And a similar fraction of households not on the lists had joined. Thus we do not exploit this further in our analysis. The idea of estimating ATT effects through this method has been attempted by Attanasio *et al.* (2011) with similar results.

The simplest estimation of impacts in an intention-to-treat framework is a comparison of means in the treatment and control

²⁵ Note that because we do not have perfect compliance, this will underestimate the impact of VSLAs.

group after the intervention, and we estimate this as a starting point. To increase the power of the test, however, we make use of the fact that we have a balanced panel—that is, information about the same households before and after the intervention in both the treatment villages and the control villages. This allows us to use a variety of estimation techniques to increase the precision of our estimates and explore the robustness of our results. Specifically, we use a simple pooled difference-in-differences OLS estimator, which compares the change in means over time across treatment villages and control villages. Second, we utilize the balanced panel in first-differences where we subtract individual baseline from endline values, which improves the precision of our estimated coefficients by removing differences in the household outcome levels at baseline.

To assess the robustness of the findings, we use a series of control variables from the baseline to perform what Freedman (2008b) calls regression adjustments, allowing for heterogeneous treatment effects across the baseline characteristics of the household. Finally, we make an attempt to estimate the average treatment of the treated using propensity score matching with first-differencing, which is our final estimation strategy. The following subsections describe each of these methodologies in turn.

Difference in Means

Difference in means is the perhaps the simplest effect estimator, as it is merely a t-test comparing averages in treatment and control villages. Due to the fact that our baseline shows balance across treatment and control, any change must be result of the intervention. Specifically, we fit the following model:

$$y_{ij} = \alpha + \beta Z_j + \theta Block_j + \varepsilon_{ij} \quad (1)$$

where y_{ij} is the outcome of household i in village j and Z_j is a dummy indicating whether village j was assigned to treatment. Following the recommendation by Duflo *et al.* (2007), we include

$Block_j$, which is a vector of indicators for the seven blocks we used to stratify the random assignment of villages.

Difference-in-Differences on Pooled Data

As already mentioned, the panel data allows us to increase precision of the estimate, thereby increasing the power of the experiment. One commonly applied estimator in RCTs with panel data is the difference-in-differences estimator on pooled data. This basically compares the change in mean outcomes for the treatment village over the time period with the equivalent change in the control villages. Hence we estimate the ITT as the difference in the differences. We fit a model often used to analyze panel data in randomized control trials (Angrist and Pischke 2009, McKenzie 2012):

$$y_{ijt} = \alpha + \beta Z_j + \gamma I_t^{2011} + \delta^{DiD} (Z \cdot I^{2011})_{jt} + \theta Block_j + \varepsilon_{ijt} \quad (2)$$

where y_{ijt} is the outcome of household i in village j at time t , Z_j is still a dummy indicating whether village j was assigned to treatment which is constant over time, I_t^{2011} indicates whether the observation is from the 2011 survey and $(Z \cdot I^{2011})_{jt}$ is the interaction of the two—i.e. a dummy that equals one if the observations is from a treatment village at the 2011 survey, zero otherwise. Hence δ^{DiD} is our parameter of interest—the ITT using pooled difference-in-differences.²⁶

As described in the section on weights above, we use the inverse sampling probability as the weight for each household in the above regression. Since the intervention is randomized at the village level, and since people in the same village are likely to have correlated outcomes, we cluster the standard errors at the village level, as recommended by Duflo *et al.* (2007).

²⁶ It is worth noting, that (1) effectively estimates the ITT as the difference between the average change in the outcome variable in treatment villages and control villages.

First-Differencing

We observe the same households both prior to and after implementation of the intervention, which allows us to potentially improve the efficiency of the estimator even further. Being more specific, we can take out any time-invariant unobserved heterogeneity at the household level by doing a first-differencing transformation, effectively differencing out any household-specific component (Wooldridge 2010, chapter 6).²⁷ This also differences out any other time-invariant parameters, meaning the regression of interest becomes:

$$\Delta y_{ij} = \alpha + \delta^{FD} \Delta(Z \cdot I^{2011})_j + \Delta \varepsilon_{ij} \quad (3)$$

Here, δ^{FD} is the first-difference estimator. Effectively, equation (2) from the previous section implies that we first average the outcomes for the treatment and control groups in each time period and then estimate the ITT as the difference in the differences of these means across time. Using (3) we instead compute the difference in observed outcomes for each household, before estimating the ITT as the differences in the means across treatment and control villages of these changes in household outcome indicated by δ^{FD} in (3). If the outcome in period one predicts the outcome in period two to a significant extent, the precision of (3) should be higher than (2). We estimate both versions for all results since the former is the standard setup, whereas the latter exploits more information in the data.

Adjusted Regressions

The difference-in-differences and the first-difference estimations are the primary regression models used below, since they depend on few variables and are commonly used. Nevertheless, other methods have been suggested to increase precision even further. In the presence of

²⁷ Some authors, for example Glerup *et al.* (2010), label this estimator “difference-in-differences,” whereas others reserve this for the aggregate differences as we do above, for example Wooldridge (2010) or Angrist and Pischke (2009). We prefer the latter to clearly separate the two estimators from each other.

heterogeneous effects across different types of households, variance can be further reduced by controlling for baseline covariates. This is useful when the change in the outcome is correlated with one or more of these covariates.

Prior to randomizing, we minimized baseline imbalances by using a stratified allocation of treatment villages and control villages by dividing villages into seven blocks as described in the section on the design of the experiment above. Nevertheless, this does not necessarily ensure that control villages and treatment villages are following the same trend. Our use of first-differencing reduces variance from time-invariant characteristics through the household fixed effect, but variance from heterogeneous trends and thus heterogeneous treatment effects over time remains.

Regression adjustment in the analysis of randomized control trials using regression (that is, where explanatory covariates are included along with the treatment indicator) is often done simply to increase precision of the treatment variable as long as covariates are predefined, correlated with the outcome variable and unaffected by treatment (Duflo *et al.* 2007). However, Freedman has shown that such regression adjustment can actually hurt precision instead of improving it (Freedman 2008b, Freedman 2008a). This has been analyzed further by Lin (2013) who adds that as long as the trial design is not very imbalanced, covariate adjustment tends to increase precision.²⁸

Following Lin, we test the sensitivity of the estimates by fitting difference-in-differences models where we include a full set of covariate-treatment interactions, where the covariates have been demeaned with the full sample mean, $\bar{X}$. Conceptually, the adjustment in equation (4) below corresponds to running separate regressions of y_{ijt} on household- and village-level baseline covariates, X_{ij} in the treatment and control groups and using these to predict the average

²⁸ Lin finds that if the trial design is very imbalanced or there are heterogeneous treatment effects that are strongly correlated with covariates, then interacted adjustments (where there is a full set of treatment and covariate interactions) will tend to improve precision as long as the number of covariates is much smaller than the number of observations in the smallest trial group. Moreover, Lin shows that using the interacted adjustments cannot hurt precision.

outcome for the entire sample if everyone were assigned to treatment and control, respectively. The intention-to-treat effect is denoted δ^{Adj} in the following regression model:

$$y_{ijt} = \alpha + \beta Z_j + \gamma I_t^{2011} + \delta^{Adj}(Z \times I^{2011})_{jt} + \pi X_{ij} + \tau Z_j \times (X_{ij} - \bar{X}) + \theta Block_j + \varepsilon_{ijt} \quad (4)$$

Matching with First-Differencing

Non-compliance gives rise to a number of challenges regarding the analysis. The fact that only approximately forty percent of households in treatment areas participated, reduces the power of the test, compared to a situation of full compliance. This is because we compare all households in treatment villages to all households in control villages. Non-compliance also makes it difficult for us to estimate the average treatment effect on the treated (ATT). If we could compare individual participants to potential participants in the control group, then we could estimate the ATT. The problem is that we do not know who the potential participants are.

Matching is one way of making a qualified guess as to who these people are. Specifically we use matching to choose observations from the control villages that are similar to the participants in treatment villages. In effect, we duplicate some control households and discard others. We use a nearest-neighbor propensity score-matching strategy and first-differencing with weights. The following describes each of the components in this strategy.

The typical goal of a matching estimator is to overcome unobserved household-specific heterogeneity that might result in different expected trends for the treatment and control groups (Smith and Todd 2005, Leth-Petersen 2010). In our case, we use it for a different reason, since we expect that the trends are similar in expectation due to the randomization and in reality due to the balanced baseline. Instead, we employ matching to identify a subgroup in the control group based on a multi-dimensional covariate vector, which we would otherwise not be able to identify.

First, the propensity to participate in VSLA is estimated using a weighted probit model in the sample of households from treatment villages only, using a wide range of baseline characteristics to predict VSLA participation. We did not use a specific balancing test, but included the covariates we believe are relevant.²⁹ Using the coefficients from this model, we then extrapolated the model onto the sample of households from the control villages, predicting the participation probability of these. We checked visually for common support by examining the propensity score densities for the treatment and control groups, as recommended by Lechner (2008) and found no reason to exclude households on this ground. We then identified the nearest control village neighbor for each household from the treatment villages that participate in VSLA, based on this predicted propensity score (Rosenbaum and Rubin 1983) allowing for households from the control villages to be used as multiple matches.

Finally, we estimated the ITT as the mean difference in first-differences between the treatment households and matched control households. In doing this, we applied the sampling weights of the treatment households to each pair of treatment-control households. This use of sampling weights has an intuitive appeal. In the words of (Reynolds and DesJardins 2009, p. 77) “the sampling weight of the untreated units is irrelevant because the method selects the nearest neighbor regardless of how many observations (in the population) that nearest neighbor represents.” However, use of sampling weights in matching is still very much an underexplored topic within the econometrics literature. In practice we implement this, while still clustering standard errors at the level of the randomization, i.e. the village level, by running weighted OLS regressions on the matched sample for each of the primary outcomes.

²⁹ The covariates are: household interest in participating prior to rollout, consumption, household size, age of the household head, education of the household head, indicator for female-headed household, dependence ratio, a subjective health measure, network size, fertilizer use on maize, number of income-generating activities, number of group memberships as well as indicators for casual labor supply, fishing, farming, owning a paraffin lamp, having sand on the floor, owning a water pump, and renting land. Finally, we include the blocks used in stratifying the randomization.

3.5 Estimating LATE

The above strategies let us estimate the intention-to-treat effects — that is, the average effect on all households in treatment villages compared to control villages. However, it would be equally interesting to say something about the effect of the treatment for those who participate. Using the randomization as an instrument, we can estimate the average treatment effect for those households whose treatment status was affected by the randomization. In other words, we compare the outcomes of those households in the treatment villages who participated due to intervention being introduced with the households from the control villages that would have participated, had the intervention started up there. This is known in the literature as a local average treatment effect (LATE) and can be estimated using an instrumental variable (IV) approach (Angrist and Imbens 1996).

Four assumptions are needed when going from estimating ITT to LATE, following (Angrist and Pischke 2009): First stage, independence, monotonicity, and the exclusion restriction.³⁰ The three first are likely to be fulfilled in our case. First stage simply means that the instrument must affect the treatment; in our case that assignment to a treatment village must make it more likely that a household participates in VSLA. Monotonicity implies that living in a village assigned to treatment must not make it less likely that the household joins a VSLA. The existence of what is known as defiers would violate this assumption (Angrist and Pischke 2009), as they are defined by not complying with a treatment exactly because they were instructed to comply. Both of these assumptions seem reasonable in our case. Independence says that the instrument is as good as randomly assigned and that the potential outcomes are independent of the instrument. Since our instrument is the random assignment, we can safely assume independence.

³⁰ Note that Wooldridge (2010) (W) defines independence differently than Angrist and Pischke (2009) (AP), in that W's independence implies both AP's independence and exclusion restriction (assumption ATEIV.1b, pp. 937). First stage is W's assumption ATEIV.1c. Monotonicity, however, is defined similarly by W and AP.

The exclusion restriction is an extension of the independence assumption in that it requires that the only way the instrument can affect the outcome is through the treatment. In our case, this means, for example, that the non-compliers cannot be affected by the compliers. In other words, there must be no spillover effects from the compliers in the treatment villages to the non-compliers in the same treatment villages.³¹ If, for example, the introduction of VSLAs improves the existing local financial market to the extent that even the households that are not members of a VSLA now have better access to credit opportunities, this will provide an upward bias on the LATE estimator. The IV-estimation attributes the entire effect to the participants, but in fact it is driven jointly by participants and non-participants. The real LATE is then lower than our estimated late, but the intervention has effects among nonparticipants as well. Unfortunately, it is not possible to test whether this is the case.

We implement a two-stage least squares instrumental variables approach on first-differenced outcomes for estimating the LATE. Using the notation from above, and letting T_{ijt} indicate whether a household actually participated in a VSLA, we first estimate the probability of a household participating in a VSLA using a linear model:

$$\Delta T_{ij} = \alpha + \beta \Delta(Z \cdot I^{2011})_j + \Delta e_{ij} \quad (5)$$

Here, the result of the randomization for village j , Z_j , should make it more likely that a household from a treatment village, $Z_j = 1$, participates in the VSLA. Using the predicted value, $\hat{T}_{ijt}$, from this first-stage, we subsequently estimate the LATE, using first-differenced outcomes as the dependent variable:

³¹ This type of spillover has a similar effect on the LATE estimator, as spillover effects from treatment to control villages would have for estimating the ITT effect. Here there would be a downward bias, though, because the observed difference between T and C villages gets smaller.

$$\Delta y_{ij} = \alpha + \delta^{IV} \Delta T_{ij} + \Delta \varepsilon_{ij} \quad (6)$$

In effect, the LATE assigns the change in outcome variables to the share of the households in treatment villages induced to participate through the randomization. Equation 5 above estimates the difference in VSLA-participation across treatment and control villages. ΔT_{ij} is identical for all households in the treatment and control groups respectively, with the $\widehat{\Delta T}_{ij}$ being equal to the change in VSLA-participation between 2009 and 2011. This estimated difference is then used to explain the differences in outcomes between treatment and control villages using equation 6.

The estimated LATE is closely related to the more standard notion of the average treatment effect on the treated (ATT). The ATT would in our case be the average effect of participating in a VSLA. However, since we have participants in both treatment and control villages—in other words, we have noncompliance with the randomization in the control villages—our LATE is not the ATT. This would be the case only if no household from the control villages had participated.

4. Results on Predefined Outcomes

Following the sections above on our estimation strategy and the gains from relying on predefined outcomes, the current section reports the results of our primary analysis. The first three columns of table 7 below present the estimated intention-to-treat effects of the VSLA intervention using difference in means, pooled difference-in-differences, and first-differenced regressions methods on the predefined outcomes. The predefined outcomes are arranged in the three groups as suggested by the LFA: increased food security, increase in income-generating activities, and improved household income.

When we use the simple difference in means approach, three variables are significant. First of all, VSLA savings increase dramatically. In other words, it seems the intervention had an effect on household decisions so that the households from the treatment villages

used the VSLAs to save. At the same time, there is a statistically significant increase in total savings, which includes all highly liquid savings. Furthermore, income-generating activities and land holdings decreased. The difference in means for land ownership is likely to be a result of imbalance at baseline, where the control group owned significantly more land than the treatment group.

A decrease in the number of income-generating activities could be explained by specialization. While access to credit might spur the startup of more enterprises, it might also cause households to specialize in one or more of their existing enterprises, thereby having a negative effect on the total number of IGAs. Similarly, increased specialization could also be the result of more efficient ex-post-risk coping: Knowing that households can rely on VSLAs might cause them to concentrate their investments and efforts into one income-generating activity, thereby taking a higher risk. As we shall see, however, this effect disappears once we move on to the more precise estimators. This is consistent with the baseline situation, where treatment households had fewer income generating activities than treatment households. Although not significant, the t-value of this difference was -1.66.

There is a clear gain in efficiency when we make use of the baseline data and take into account either baseline means in the difference-in-differences estimation or household-specific time-invariant effects using first-differencing transformation. This underlines the importance of multiple observations over time, which was not taken into account in the earlier estimated minimum detectable effects (see section 2.3). Except for land holdings, all signs are the same for the three specifications, and the point estimates do not differ much. The standard errors decrease across the board as we move from difference in means to pooled difference-in-differences to first-difference. This causes a number of the estimated ITT-effects to become significant.

Table 7. Effect on Predefined Outcomes

			(1) Difference in Means	(2) Difference-in- Difference	(3) First- Difference	(4) IV	(5) Adjusted Regressions	(6) NN Matching
<i>Increased Food Security</i>								
Number of months with fewer than three meals a day	3422		-0.205 (0.38)	-0.501 (0.45)	-0.571 (0.38)	-2.168 (1.40)	-0.374 (0.31)	-0.401 (0.53)
Number of meals yesterday	3422		0.055 (0.04)	0.147** (0.07)	0.128** (0.05)	0.487** (0.20)	0.071** (0.03)	0.014 (0.06)
Total food consumption per week per adult equivalent (MK, log)	3422		0.023 (0.05)	0.079 (0.07)	0.093 (0.06)	0.351 (0.23)	0.101* (0.06)	0.045 (0.08)
<i>Increase in Income Generating Activities</i>								
Number of income-generating activities (including agriculture and livestock)	3422		-0.175*** (0.06)	-0.105 (0.11)	-0.056 (0.07)	-0.213 (0.26)	-0.053 (0.07)	0.147 (0.13)
Total savings (log)	1625		1.384*** (0.51)	0.969 (0.90)	0.948 (0.79)	2.974 (2.20)	0.577 (0.48)	0.894 (0.54)
VSLA savings (log)	1625		2.492*** (0.58)	2.413*** (0.66)	2.393*** (0.57)	7.505*** (0.63)	1.926*** (0.36)	4.458*** (0.53)
<i>Improved Household Income</i>								
USAID PAT per capita consumption (log)	3422		0.03 (0.03)	0.041* (0.02)	0.042** (0.02)	0.159** (0.08)	0.029* (0.02)	0.024 (0.03)
Size of house (number of rooms)	3422		0.104 (0.10)	0.133* (0.07)	0.154** (0.06)	0.584** (0.26)	0.134** (0.06)	0.192* (0.11)
House has cement floor	3422		0.019 (0.02)	0.001 (0.02)	-0.011 (0.02)	-0.041 (0.07)	-0.009 (0.02)	0.010 (0.03)
Land ownership (acres)	3422		-0.272* (0.14)	0.11 (0.16)	0.179 (0.14)	0.681 (0.52)	0.08 (0.13)	0.242 (0.18)

The table shows effects on food security and household income. Column (1) ignores baseline information, whereas column (2) controls for the baseline value of the outcome. Column (3) improves precision by using a first-differencing transformation of data. All specifications include dummies for stratification blocks (not reported). Regressions are run on a sample where none of the outcome variable have missing values and number of observations are for the difference-in-differences estimation. Standard errors in parentheses, clustered at the village level. * p<0.10, ** p<0.05, *** p<0.01.

For the difference-in-differences estimation on the pooled data, three outcome variables are positive and significant. VSLA savings is significant at the 1% level, but particularly, the number of meals consumed yesterday³² increases: the households from the treatment villages consume on average 0.147 more meals a day ($p < 0.05$). Given that the baseline average is 2.65, this is sizable. There are also positive effects on consumption when measured by the USAID PAT as well as the number of rooms. The logged value of consumption increases by 0.041 – i.e. a 3% increase ($p < 0.1$).³³ Number of rooms increases by 0.15 and is significant at the 10% level. The latter might be surprising, but given the rural setting, where households usually live in huts built of mud on a wooden frame, it is quite common to build additional rooms onto the existing structure when money allows, or—if possible—to replace the mud hut with burnt or un-burnt brick structures. We find no significant effects, however, on the type of floor in the dwelling. We also find null results (i.e., no significant effects) on food consumption, length of the hungry period, number of income-generating activities, and total savings. Compared to the difference in means, land ownership changes sign and loses significance, when we take the baseline difference into account.

When we move to the first-difference estimation, results are similar to the difference-in-differences on pooled data. Treatment households increased their VSLA savings. Number of meals yesterday and number of rooms are still significant at a 5%, but so is consumption measured by USAID PAT and number of rooms. There are still no results on food consumption, length of the hungry period, number of income-generating activities, and total savings.

As for the null results, a number of explanations can be given. Both the hungry period and food consumption are difficult to measure. The former is measured using a year recall—that is, asking households about their average consumption in meals per day during

³² i.e., the day before the interview was conducted.

³³ Following Kennedy (1981), we interpret the implied effects of the coefficients using the formula

$$e^{\left(\hat{\beta} - \frac{\text{var}(\hat{\beta})}{2}\right)} - 1.$$

all months of the past year. Recall periods of more than one week are known to be particularly biased (Deaton 2001). Regarding consumption, our baseline survey mirrors the consumption module of the Second Integrated Household Survey, which we found was prone to error, in particular regarding the quantity measurements. Our consumption measure and its limitations are documented in Appendix A. Total savings, which were also insignificant, can be stable if VSLA simply changes the savings pattern, something we return to below. Finally, regarding the null effect on income generating activities, effect on consumption and rooms can happen through other channels than income-generating activities, something we also elaborate on below.

4.1 Local Average Treatment Effects

As described in the section on empirical strategies, we can use the randomization as an instrument to extract the local average treatment effect (LATE). Rather than estimating the average effect of assigning a village to treatment for all households in the treatment village, the LATE is the estimated effect of actually participating for those households that were induced into participation by the randomization—i.e., for the compliers from the treatment villages. The fourth column in table 7 displays the results of the two-stage least squares estimation of the LATE.

Estimating the LATE has two hardly surprising consequences in general when compared to the ITT estimates: The estimated coefficient increases as the effect on the households in treatment villages—which the ITT estimator assigns to all households equally (even those that did not participate)—is averaged by LATE over the households that actually participated. A simple way of understanding the substantial increase in the estimated effect is by dividing the estimated coefficient from the second column by the additional proportion of households that participated in the VSLA in the treatment villages. This was 23.2%, so roughly a quarter. The IV estimate is approximately four times larger than the intention-to-treat

effect, which divides the treatment effect (enjoyed by VSLA participants) by all households in the village.

At the same time, the standard errors also increase, since we introduce more noise by using the estimated probability of being a VSLA participant rather than the actual assignment to treatment. The LATE estimates are significant for the same outcomes as the first-difference ITT, and the estimated effects are quite substantial: participants ate almost half a meal more than nonparticipants the day before the survey, and their consumption increased by thirteen percent using PAT. Furthermore, their VSLA-savings savings increased, and they expanded their dwelling with half a room more as a consequence of the intervention.

Whereas these figures seem high, it is important to remember that the LATE is not the average treatment of the treated. If, for example, there is spillover from participants in treatment villages to non-participants in the same villages, then the real average treatment of the treated is lower than the LATE we find here.

4.2 Robustness

We take a number of steps to assess the robustness of the results found above, focusing on the estimated intention-to-treat effects. First we use adjusted regressions to remove variance from heterogeneous effects as explained in the section on the empirical strategy above. Second, we make an attempt to make up for the fact that there is non-compliance by using matching with first-differencing. The results are in the two last columns of table 7. Second, we assess how sensitive the results are to the choice made regarding weights, split households, and the assignment of interviewers across treatment and control villages. These robustness checks are presented in table 8. Finally, we discuss Hawthorne and John Henry effects and their potential implications for the results. Throughout this section we compare results to the results using first-differencing, since we believe that the individual fixed effects provide the highest precision possible with the current data.

Alternative Specifications: Adjusted Regressions and Matching

Columns 5 and 6 of table 7 report the results from using the adjusted regressions, as well as from an estimation using nearest-neighbor propensity score matching on first-differenced outcomes. The results using adjusted regressions are very similar in size and significance to the results from the first-differences estimator.

There are some differences, however, when looking at the matching estimates, not so much regarding the direction of the effect, but more regarding the (lack of) significance. Two variables are still significant and positive, namely savings in VSLA and the number of rooms. The rest are insignificant. This could be due to several factors. First, the propensity score does not perfectly predict participation. In the treatment group, sixty-seven percent are correctly predicted; that is, they have either a propensity score below 0.5 and are not participants, or they have a propensity score above 0.5 and participate in a VSLA. Second, matching requires a number of covariates for estimating the propensity score, which decreases the sample size due to missing values, and thereby also lowers the precision of the estimated effects. Third, the estimation uses sampling weights from the treatment observations for both the treatment and control household, as described in the section on the empirical strategy. In the literature, sampling weights are not commonly applied to propensity score matching, and the results are quite sensitive to the use of these weights.

Sampling Weights

While we do believe the sampling weights should be applied in order to estimate the parameters of interest for the entire population in the survey area, it is nevertheless of interest to know how sensitive the results are to the sampling weights. In other words this will provide some indication of whether the sampled population differs from the population in the survey area.

Column two of table 8 shows the un-weighted results using the first-differences approach. The sampling strategy of oversampling households that had expressed an interest in the VSLA intervention

carries over to the weights. Although results are similar to a large extent, there are enough differences to say that our survey population differs from the entire population in the area: Some of the coefficient changes and the significance levels are affected. The estimated ITT effect on the number of meals yesterday and USAID PAT diminishes and is no longer significant. At the same time, total savings are now significant. Number of rooms is significant, albeit only at a ten percent level.

Table 8. Importance of Sample Weights, Split Households and Interview Fixed Effects

Outcome		First-Difference	First-Differences Unweighed	First-Differences Without Split Households	First-Differences, Interviewer FE
Increased Food Security					
Number of months with fewer than three meals a day	3422	-0.571 (0.38)	-0.288 (0.34)	-0.547 (0.38)	-0.419 (0.34)
Number of meals yesterday	3422	0.128** (0.05)	0.036 (0.04)	0.131** (0.05)	0.092** (0.04)
Total food consumption per week per adult equivalent (MK, log)	3422	0.093 (0.06)	0.078 (0.06)	0.092 (0.06)	-0.002 (0.05)
Increase in Income Generating Activities					
Number of income-generating activities (including agriculture and livestock)	3422	-0.056 (0.07)	-0.059 (0.07)	-0.069 (0.07)	0.09 (0.07)
Total savings (log)	1625	0.948 (0.79)	0.713* (0.39)	1.01 (0.80)	0.193 (0.44)
VSLA savings (log)	1625	2.393*** (0.57)	2.275*** (0.36)	2.431*** (0.57)	2.133*** (0.51)
Improved Household Income					
USAID PAT per capita consumption (log)	3422	0.042** (0.02)	0.027 (0.02)	0.040* (0.02)	0.023 (0.02)
Size of house (number of rooms)	3422	0.154** (0.06)	0.125* (0.07)	0.159** (0.06)	0.091 (0.07)
House has cement floor	3422	-0.011 (0.02)	-0.008 (0.02)	-0.01 (0.02)	0.005 (0.02)
Land ownership (acres)	3422	0.179 (0.14)	0.018 (0.15)	0.195 (0.15)	0.113 (0.12)

Note: The table shows that the main results are robust to the exclusion of split households, and consistent with results with interviewer fixed-effects included. All specifications include dummies for stratification blocks (not reported). Regressions are run on the same sample as the main results regressions. Standard errors in parentheses, clustered at the village level. * p<0.10, ** p<0.05, *** p<0.01.

Importance of Split Households and Interviewers

In this section we investigate whether the results could be driven by some of the choices made during the analysis and collection of data. We focus on two of these choices, which we believe are the most important. One is how to deal with households that had split up during the time period between the two surveys. The second issue is recognizing that the data very much reflects the ability of the interviewers to gather the correct information from the respondents.

During the 2011 survey, we took care to track any households that had moved or split up during the survey period. Thus, for a number of households, defined from the 2009 survey, we had two observations in the 2011 survey. If, for example, the household head and spouse were interviewed as part of the same household in 2009 and got divorced in the following two-year period, they were each interviewed as being part of their new households in 2011. In total we have 54 such split households in the 2011 survey, meaning twenty-seven of the households from the 2009 survey had split up two years later. In the primary analysis above, we simply randomly dropped one of each of these split households from the 2011 sample. In the third column of table 8, we estimate the ITT using the first-differences estimator, but we drop all these split households. The point estimates do not change much; neither do the significance levels compared to the first-difference estimator, where we include one randomly selected split household. The only difference is that USAID PAT is significant at a 10% level, not 5%. We conclude that the split households do not matter for the conclusions.

In the fourth column we take the interviewer effects into account by including a full set of dummies for the interviewer of both the 2009 and 2011 survey. This will take into account any differences in levels across interviewers. Of course, the interviewers might affect the data beyond the mere levels, but this is, to our knowledge, the best we can do. Including interviewer fixed effects does change the estimates as well as the significance levels of the estimated ITT effects. Specifically, only meals per day is now significant. This might also be

due to the fact that this estimation requires estimating an additional fifty-two parameters—one for each of the interviewers used in 2009 or 2011—although these are not reported, and thus a corresponding loss in degrees of freedom.

Placebo, Hawthorne, and John Henry Effects

Since the present study was not blinded in any way, we risk various biases including placebo, Hawthorne, and John Henry effects. Placebo effects are possible if an effect is due to the attention and processes given by the project to the treatment group, but not the actual treatment. There is no doubt that the VSLA intervention being analyzed is much more than financial services, for example strengthening of social ties, exchange of knowledge and increased spreading of norms, as described in the section on expected impacts of the intervention. Any effect through these channels is impossible to disentangle from the effect of the actual financial service. The consequence is that what we evaluate is the entire VSLA package, not just the effects of improved access to savings and credit.

Hawthorne effects are the effects of being studied. We might observe this effect in two areas: on participants in the study and on the implementing agency and field officers. The former is unlikely to be important as it applies to both treatment participants and control participants. Importantly, there was no overlap of interviewers and staff working in the villages. Moreover, interviewers were carefully instructed not to mention any linkage between the two, and the survey was presented as "Karonga Vulnerability Study." The latter, however, might matter for external validity: It is possible that the field officers implemented the VSLA more thoroughly knowing that a study was ongoing. As such, we might not be estimating the effect of the VSLA package, but the effect of the VSLA package when implementers are being monitored closely.

Finally, John Henry effects, which refer to specific effects on the control group due to the fact that they participate in an experiment, might be relevant. The intervention started in control areas after the endline survey and all participants were made aware at the time of the

awareness meetings in 2009 that the implementing organization would roll out the intervention in their area in the coming three years. Knowing that VSLAs would start in their village sometime in the future, households might have changed their behavior and postponed investments, started similar activities themselves, or become upset with the concept and the implementers because they were forced to wait. This could result in both positive and negative bias in the estimated impacts reported above.

5. Intermediate Outcomes

The previous section identified specific impacts of the intervention on some of the predefined outcomes, specifically number of meals yesterday, total consumption using USAID PAT, VSLA savings, and number of rooms. Are these effects and their size justified by the intervention? In the section on expected impacts, we hypothesized about some of the channels through which the intervention might affect household welfare. While we did not propose a theoretical framework with testable implications, the current section nevertheless tries to provide some indications of whether some of these channels can explain the observed effects on household welfare.

We started out by investigating whether the intervention actually improved access to and take-up of the primary financial services offered: savings and credit. Given the self-reported use of the money from the share-out and loans taken from the VSLAs, we followed this stream of money. Households reported using share-outs and loans for increased investments in agriculture and other household-owned small scale businesses, and we therefore investigated the household's agricultural production as well as the evidence on other income-generating activities.

5.1 Savings Volume and Share-Out

The core component of the VSLA intervention is savings. Table 7 therefore shows the estimated ITT-effects on the total savings volume and savings in the VSLA intervention. Table 9 below extends this analysis by looking at the effects on other sources of highly liquid

savings. Column one lists the baseline means, whereas column two and three contain results from regressions on pooled data and differenced outcomes, respectively. Apart from the result on savings in the VSLA, we do not observe any significant changes in the other savings options.

A natural question is where the funds saved in the VSLA come from, since the other savings options do not decline. Another interesting issue is why total savings do not increase. Several explanations are possible. The non-VSLA savings actually do decline, although they are not statistically significant, while total savings increase a little. So the VSLA savings might come in part from the other savings options, and in part from extra income, whether this is profits from a business or sale of assets. The only certain conclusion seems to be that households now use the VSLA option.

Table 9. ITT-Effects on Savings Outcomes

Outcome	Baseline Mean	N	Difference- in- Difference	First- Difference
Total savings (log)	5.662 [3.811]	1637	0.677 (0.91)	0.712 (0.77)
VSLA savings (log)	0.17 [1.163]	1625	2.413*** (0.66)	2.393*** (0.57)
Non-VSLA savings (log)	5.583 [3.832]	1645	-0.536 (1.00)	-0.367 (0.70)
Savings with friend/relative (log)	0.524 [2.104]	1637	-0.037 (0.43)	-0.086 (0.39)
Savings at home (log)	4.826 [3.816]	1637	-0.792 (0.89)	-0.677 (0.62)
Savings with bank (log)	0.767 [2.642]	1636	0.361 (0.53)	0.544 (0.45)

Note: The table shows that the intervention increased savings in VSLAs, but not enough to cause an increase in the total level of savings. Standard errors in parentheses, clustered at the village level. Standard deviations are in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dummies for stratification blocks included in the regression, but not reported.

When further exploring the use of VSLA, it is important to note that saving in the groups is not identical to saving in a regular savings account: a core feature of the intervention is the annual share-out of all

savings along with any interests earned from loans made during the cycle.

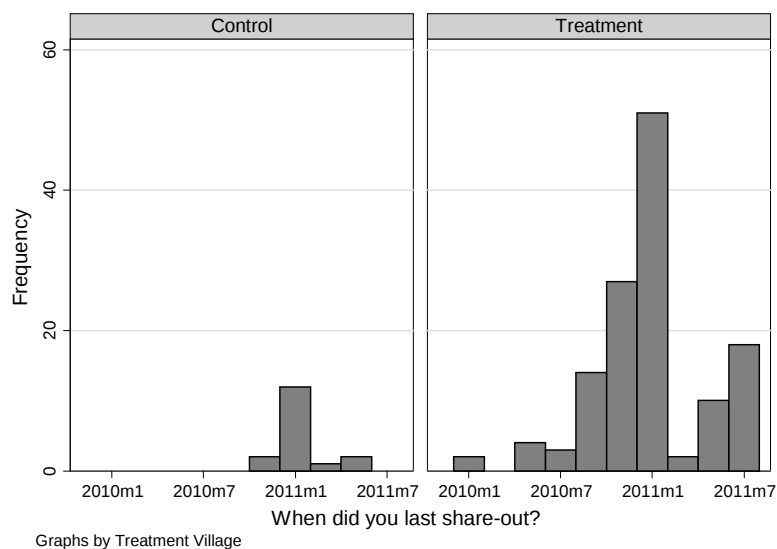
Understanding the impacts of the intervention necessitates understanding how this money from the share-out is spent. Table 10 below shows the self-reported primary use of the money from the share-out. The first column shows the estimated number of households reporting the different uses, and the second column the percentage of households that have shared out. While we estimate that thirty-nine percent of the households in the treatment villages had joined the VSLAs (table 6), only sixteen percent of the households had actually shared-out at the time of the survey in 2011 (table 10).³⁴ Among these households that have shared out, forty-four percent report using the money primarily for agricultural inputs or investments. This corresponds well to the timing of the share-out observed in our area, which was typically around the time when the planting of new crops was carried out (figure 5).

³⁴ From the information recorded by the implementing NGO, only 3 of the 102 groups that had been initiated by September 2011 had shared-out twice, while another 40 groups had shared-out once.

Table 10. Self-Reported Primary Use of Money from Share-Out

Use	Number	Percentage
Agricultural inputs	186	30%
Agricultural investments	85	14%
Buy livestock	15	2%
Trading	89	15%
Education	15	3%
Health	14	2%
Ceremonies	0	0%
Food consumption	61	10%
Emergency	5	1%
Items for the household	79	13%
Other	63	10%
Total share-outs	613	100%
Observations used in estimating totals	130	
Percentage who have shared-out	16%	

Note: The table shows that the primary use of share-out was agricultural inputs and investments. Totals and percentages are estimated using sampling weights.

Figure 5. Timing of Share-Outs

5.2 Credit Volume

The other key component of the intervention is the use of pooled savings as credit for the members of the group. Table 11 below shows the estimated ITT effects on a range of credit-related outcomes using our pooled difference-in-differences as well as the first-difference strategy. Just as the savings component of the intervention was used by households in the treatment villages, so was the credit component. Living in a treatment village increased the number as well as the value of loans active within the past twelve months (total loan amount). The intervention also increased the take-up of loans for investment purposes: The probability of having taken a loan for investment purposes increases; the number of loans taken for investment purposes also increases; and so does the value of loans taken for investment in agriculture.

Table 11. ITT Effects on Credit Outcomes

Outcome	Baseline Mean	N	Difference-in-Difference	First-Difference
Household had any loan in past 12 months	0.07 [0.26]	832	0.134** (0.06)	0.142*** (0.05)
Household took out loan for investment purposes in past 12 months	0.044 [0.21]	832	0.075* (0.04)	0.079* (0.04)
Number of loans active within past 12 months	0.073 [0.28]	832	0.148** (0.06)	0.157*** (0.05)
Total loan amount (log)	0.606 [2.24]	832	1.267** (0.51)	1.338*** (0.43)
Number of investment loans	0.046 [0.22]	832	0.084* (0.04)	0.088** (0.04)
Total amount borrowed for agricultural investments (log)	0.25 [1.52]	832	0.549* (0.29)	0.546** (0.22)
Total amount borrowed for business purposes (log)	0.148 [1.12]	832	0.153 (0.26)	0.183 (0.24)

Note: The table shows that the intervention caused an increase in household borrowing across various types of loans and different specifications. Standard errors in parentheses, clustered at the village level. Standard deviation in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dummies for stratification blocks included in the regression, but not reported.

The increases are quite dramatic. Total loan amounts increase by 216%, and increase in the amount borrowed for agriculture is similarly larger (48%), both using the first-differencing.³⁵ The figures are significant at a 1% and 5% level, respectively. The agricultural loans are somewhat surprising given a short loan period of three months. We can think of two reasons. Certain crops have faster growth cycles, for example ground nuts and tomatoes. But it is also possible that the loans were indeed for longer term agriculture, but that low liquidity made repayment over time easier.

The background for these huge changes is the very modest starting points, however. Loan values were almost zero at the time of baseline and so was the share of households with loans. If we count all loans, the share of households with loans tripled from 7% to 21%. We conclude that the intervention successfully increased borrowing in the area.

While table 11 does indicate that the total number of loans increased, table 12 below shows the self-reported use of any loans obtained from the VSLA in the past 12 months at the time of the survey in 2011. Within both treatment and control villages, the first column shows the estimated number of loans taken, and the second column shows the percentage of total loans. Here, agricultural investments and inputs do not play as predominant a role: nineteen percent report this as the primary use of the loan. This might be due to the relative short nature of the loan from the VSLA, which usually has to be paid back within three months. Instead, trading and business is the use of ninety-four percent of all loans. Based on focus-group interviews, we did indeed hear of VSLA group members who took out a loan, bought rice in the rice fields located at some distance, transported the rice back to their own village, and sold it there at a profit.

³⁵ As above, we use Kennedy (1981) to get from the log point estimates to percentage change:

$$e^{\left(\hat{\beta} - \frac{var(\hat{\beta})}{2}\right)} - 1.$$

Table 12. Self-Reported Use of Credit

	Treatment 2011		Control 2011	
	Number	Percentage	Number	Percentage
Agricultural inputs	76	8%	1	0%
Agricultural investment	104	11%	4	0%
Buy livestock	4	0%	0	0%
Trading/business	311	34%	118	13%
Education	15	2%	0	0%
Health	22	2%	5	1%
Ceremonies	0	0%	16	2%
Food consumption	50	5%	4	0%
Emergency	21	2%	5	1%
Household items	149	16%	3	0%
Other	9	1%	0	0%
No loan	0	0%	0	0%
Total	761	83%	156	17%
Observations used in estimating totals	113		22	

Note: The table shows that credit is used primarily for trading and agriculture. Totals and percentages are estimated using sampling weights. The percentage reports the share of total loans.

5.3 Agriculture

Two facts caused us to investigate the effect of the intervention on the agricultural activity in the area: First, the reported use of share-outs and credit involved a large share of agricultural investments. Second, the timing of the share-out corresponded to the time of the agricultural season when large-scale investments into agricultural production were needed, for example fertilizer and seeds.

Table 13 and 14 below therefore show the ITT effects on a range of agricultural inputs as well as outputs. The first column shows the baseline mean of the outcome variables in the control villages, allowing for an interpretation of the estimated size of the ITT effects throughout the table. For a number of the measures, we only have information from the long questionnaires, which halves the sample size used for assessing the impact. The second column therefore shows the number of observations used in the pooled difference-in-differences estimation.

Relying on the difference-in-differences as well as first-differences estimators, there seem to be significant effects, although they are few and some are significant at the 10% level only. On the input side, households from treatment villages are more likely to use fertilizer when growing maize and the first-difference estimator show a positive result with regard to irrigation, something particularly necessary together with fertilizer.

Looking at the agricultural output, we are unable to detect an increase in total maize production, but there is a significant effect on selling maize as well as the value of the sale.³⁶ In the first-difference estimation these results are significant at a 1% level. One possibility is that the quality of maize increases. This evidence, although suggestive, points to a change in agricultural practices: Treatment households seem to invest more in growing their primary crop, maize, and by doing this they also get a higher output.

When we compare these results to the potential channels of impact listed earlier, this suggests that the intervention was successful in encouraging households to save up for specific investments. Through the intervention, households might have committed themselves to saving more than they would have otherwise, which allows them to invest more in agriculture subsequently. Alternatively, having the share-out at the time when agricultural investments are needed might have served as a further encouragement for additional savings. Or the implicit and explicit insurance mechanisms built into the intervention through actual insurance and access to credit and informal networks might have caused households to use ex-post rather than ex-ante risk-coping strategies. This would be the case if households also increased their use of hybrid maize, something we do not find supported in the data. In general, we are unable to assess whether one or more of these channels drove the significant changes observed in the agricultural production

³⁶ We have also looked at the productivity, defined as output (kg) per acre, and found similar although insignificant results.

Table 13. ITT-Effects on Agricultural Input

Outcome	Baseline Mean	N	Difference in Means	Difference- in-Difference	First- Difference
Household uses any fertilizer	0.689 [0.463]	1616	0.052 (0.05)	0.068 (0.08)	0.066 (0.07)
Household uses any fertilizer on maize	0.449 [0.498]	3429	0.132** (0.05)	0.107* (0.06)	0.093* (0.05)
Household cultivates vegetable plot	0.248 [0.432]	3446	-0.046 (0.03)	0.036 (0.04)	0.034 (0.03)
Household has any irrigated plots	0.25 [0.434]	1637	-0.004 (0.05)	0.078 (0.06)	0.106** (0.05)
Household used any purchased seeds	0.775 [0.418]	1616	-0.006 (0.05)	0.015 (0.05)	0.025 (0.06)
Total area cultivated (acres)	2.865 [1.728]	1616	-0.262 (0.19)	0.012 (0.22)	0.016 (0.21)
Area with maize (acres)	1.369 [0.9]	3429	-0.310* (0.17)	-0.109 (0.14)	-0.149 (0.16)
Area with local maize (acres)	0.527 [0.783]	3429	-0.306** (0.15)	-0.174 (0.12)	-0.219 (0.13)
Area with composite maize (acres)	0.079 [0.359]	3429	-0.009 (0.01)	-0.02 (0.03)	-0.019 (0.03)
Area with hybrid maize (acres)	0.762 [0.858]	3429	0.005 (0.08)	0.086 (0.08)	0.09 (0.07)
Area with tobacco (acres)	0.183 [0.438]	1616	0.021 (0.06)	0.064 (0.06)	0.067 (0.06)
Area with cotton (acres)	0.315 [0.642]	1616	-0.025 (0.06)	0 (0.07)	-0.02 (0.07)
Area with rice (acres)	0.313 [0.591]	1616	-0.181 (0.13)	-0.035 (0.11)	-0.022 (0.11)

Note: The table shows signs of a change in agricultural practices. In particular, households use more fertilizer. N is the number of observations for the regression on the pooled data. Standard errors in parentheses, clustered at the village level. Standard deviations are in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dummies for stratification blocks included in the regression, but not reported.

Table 14. ITT-Effects on Agricultural Outputs

Outcome	Baseline Mean	N	Difference in Means	Difference-in-Difference	First-Difference
Quantity of maize harvested (kg, log)	5.471 [1.545]	3429	-0.12 (0.15)	0.115 (0.15)	0.061 (0.14)
Local maize harvested (kg, log)	2.328 [2.741]	3429	-0.517 (0.31)	-0.187 (0.23)	-0.28 (0.20)
Composite maize harvested (kg, log)	0.339 [1.395]	3429	0.001 (0.07)	-0.031 (0.14)	-0.04 (0.12)
Hybrid maize harvested (kg, log)	3.533 [2.849]	3429	0.275 (0.24)	0.284 (0.29)	0.321 (0.23)
Quantity of maize harvested per acre (kg, log)	5.245 [1.686]	3429	-0.005 (0.11)	0.251 (0.20)	0.174 (0.18)
Local maize harvested per acre (kg, log)	2.212 [2.686]	3429	-0.44 (0.29)	-0.122 (0.22)	-0.215 (0.20)
Composite maize harvested per acre (kg, log)	0.323 [1.339]	3429	0.013 (0.07)	-0.014 (0.14)	-0.025 (0.12)
Hybrid maize harvested per acre (kg, log)	3.381 [2.84]	3429	0.298 (0.22)	0.329 (0.29)	0.362 (0.25)
Household sold any crops	0.588 [0.493]	1616	-0.047 (0.06)	-0.012 (0.08)	0.015 (0.08)
Household sold any maize	0.201 [0.401]	1616	0.094* (0.05)	0.142** (0.07)	0.197*** (0.06)
Value of agricultural sale (MK, log)	5.355 [4.788]	1616	-0.684 (0.67)	-0.569 (0.86)	-0.237 (0.77)
Value of maize sale (MK, log)	1.505 [3.141]	1616	0.725* (0.41)	1.001* (0.57)	1.427*** (0.48)

Note: The table shows signs of a change in agricultural practices. In particular, the total value of maize went up. N is the number of observations for the regression on the pooled data. Standard errors in parentheses, clustered at the village level. Standard deviations are in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dummies for stratification blocks included in the regression, but not reported.

5.4 Income-Generating Activities

Among the other uses of the money from share-out and take-up of credit, investments into income-generating activities—including trading—stand out with forty-four percent of treatment and control respondents reporting this as their use of credit (see table 12 above). This is also the channel through which it is generally believed that microfinance should have an effect. Because of this, we investigate this channel further, even though table 7 did not show an effect of the intervention on the predefined outcome of the number of income-

generating activities. Table 15 reports the estimated ITT effects on a number of outcome measures related to income-generating activities. Similar to table 13 and 14 above, the first two columns show estimated baseline means and the sample size for the pooled difference-in-differences regression, respectively.

The first outcome variable is the predefined variable also reported in table 7. However, this variable simply counted the number of income-generating activities the household was engaged in, and included crop production, raising livestock, fishing, trading, and other small-scale enterprises. The second variable listed in the table only counts the number of actual small-scale businesses the household is engaged in.³⁷ Using this definition, there is a slight indication that the intervention increases the number of businesses. For the first-difference estimator, households from the treatment villages increase the number of businesses by the time of the endline survey compared to the households from the control villages, a result significant at a 10% level. We are, however, unable to determine whether this is due to more businesses starting up, or whether the existing businesses are more likely to survive in the treatment villages. The third variable provides an indication of whether the VSLA helped entrepreneurs starting up businesses, but, although the estimated effect is positive, it is not significant. While we do see the number of businesses increased in the treatment area, the total income from businesses did not increase. The reason we estimate the effects on business income using both the level and the log-transformed values is the influence of outliers as well as the non-negligible number of households that reported experiencing a loss from a specific business. These are simply dropped when taking log-values.

In summary, whereas we did find significant increases in investments in agriculture, the VSLAs were less successful in increasing the income from other income-generating activities and businesses. We are unable to assess whether the investments in agriculture were simply deemed more profitable by the participating households, or

³⁷ This does not include agriculture and livestock raising, for instance.

whether the investments in small-scale enterprises took longer to materialize into increased profits and income.

Table 15. ITT Effects on Income-Generating Activities

Outcome	Baseline Mean	N	Difference in means	Difference- in- Difference	First- Difference
Number of income-generating activities (including agriculture and livestock)	1.98 [1.106]	3454	-0.181*** (0.06)	-0.104 (0.11)	-0.06 (0.07)
Number of businesses (excluding agriculture and livestock)	0.92 [0.779]	1645	0.176 (0.13)	0.258 (0.19)	0.291* (0.15)
No of businesses started with VSLA	0.00 [0.033]	1546	0.038 (0.02)	0.032 (0.02)	0.036 (0.02)
Total income from all businesses	15289.58 [59466.8]	1263	8846.643 (10318.70)	-136.579 (11064.17)	-443.021 (12291.89)
Total income from all businesses (log)	8.03 [1.823]	1201	0.486 (0.39)	0.383 (0.50)	0.509 (0.54)
Total income from all businesses (no businesses = zero income)	10429.14 [49612.5]	1645	6432.549 (8797.31)	448.083 (7801.45)	-765.38 (8119.99)
Total income from all businesses (no businesses = zero income) (log)	5.37 [4.066]	1583	0.343 (0.54)	0.597 (0.70)	0.446 (0.61)
Total income from all businesses - given business in 2009	15271.69 [59434.1]	1273	7791.096 (9956.58)	-1006.884 (10633.41)	-956.436 (11815.35)
Total income from all businesses - given business in 2009 (log)	8.02 [1.844]	1211	0.538 (0.47)	0.37 (0.56)	0.585 (0.59)
Total stock of petty trade business (log)	5.32 [3.98]	629	1.145 (0.71)	-1.207 (0.95)	1.074 (1.51)
Number of non-household members employed in enterprises	0.51 [1.541]	1645	0.017 (0.16)	-0.042 (0.17)	-0.033 (0.16)
Any non-household members employed in enterprises	0.16 [0.369]	1645	0.026 (0.02)	-0.021 (0.04)	-0.009 (0.04)

Note: The table shows that there is no effect on income generating activities. Standard errors in parentheses, clustered at the village level. Standard deviations are in square brackets. N is the number of observations for the pooled regression. * p<0.10, ** p<0.05, *** p<0.01. Dummies for stratification blocks included in the regression, but not reported.

6. Cost Effectiveness

In the previous sections we estimated the benefits for households of introducing village savings and loans associations. While we find some significant and positive effects on household welfare, including the number of meals eaten, total consumption, and number of rooms, these effects do come at a cost: the cost of the intervention. In order to assess whether the VSLAs seem worthwhile, this section provides some estimates of the cost effectiveness, which will allow for comparison with other interventions that could also have affected overall household welfare and food security.

There are a number of caveats associated with the following calculations. In general, providing a full cost-benefit analysis lies beyond the scope of this paper, and in the current section we simply provide some back-of-the-envelope calculations. First, the intervention runs for three years and by that time should reach all villages in the survey area—i.e., both the treatment villages and control villages—and we estimate the ITT and LATE effects of the intervention on households from the treatment villages after two years only. In using the cost of the intervention in the cost-effectiveness calculations, we therefore have to make assumptions about the cost incurred during the first two years of implementation compared to the final year. For simplicity, we simply estimate a cost per year of implementation as total costs divided by three. Second, the specific intervention investigated is quite costly compared to the average VSLA intervention. This is both due to the relatively small scale of the intervention, which does not exploit economies of scale fully, and the fact that VSLAs often use agents to a greater extent than the randomization allowed in the current setup.³⁸

We estimated the cost of implementing the project per household in the treatment villages and compared this to the estimated ITT effects. The total budget for the three-year intervention was

³⁸ Normally, VSLAs are introduced to a single village, where some VSLA members become village agents, who subsequently start VSLAs in the neighbouring villages. While this limits the costs of the intervention, we do not know what the effects of relying on village agents rather than trained field officers would be.

approximately USD \$201,000,³⁹ thus the cost for the first two years of implementation was approximately USD \$134,000. The implementing partner reported a total number of 1,783 members in the VSLAs by September 2011 out of the 3,800 households in the treatment villages. In other words, the cost per member was approximately USD \$75,⁴⁰ while the cost per household from a treatment village was approximately USD \$35.

Looking at the benefit side of the calculations, we face the problem of not knowing whether the estimated effects are permanent or whether they are only temporary. Using the estimated effect on total household consumption using USAID's PAT methodology, we found an increase of 3.3% for all households in the treatment villages using the first-difference estimator as reported in table 7. At the baseline, consumption per household member per day using USAID PAT was MK 75.62, corresponding to USD \$0.83. The average household consisted of 5.77 members, resulting in a total household income of USD \$4.81 per day. An increase of 3.2% thus corresponds to an increase in the average household income of USD \$0.16 per household per day from the introduction of the VSLAs. If this effect is only for one day, the increase in household income comes at quite a high cost—the USD \$35 estimated above. However, since no external funds are injected into the groups, and the intervention only has an effect through more efficient use of existing funds from the households, there is no reason to assume the effect should last for one day only. If we assume the effect lasts for a month, the cost of implementing the project results in an increase in household income of USD \$4.85. If the effects should last for a year, the benefits amount to USD \$58.04 per household from the treatment village. This is quite impressive, given the estimated cost of implementation of USD \$35 per household. Furthermore, the introduction of VSLAs may have

³⁹ Budgeted total cost was 1,130,226 DKK. Using an exchange rate of \$1 USD = 5.60 DKK, this is equivalent to USD \$201,180.

⁴⁰ This is quite high compared to the typical cost of implementing VSLAs as reported by Allen and Panetta (2010), which was USD \$18 to \$48.

affected other measures above and beyond their impact on household consumption as measured by PAT.

As noted, these back-of-the-envelope calculations come with a number of caveats. We do not know the details of the complex dynamics in the local area which drive the results. Also, the 95% confidence interval is USD \$2.7–USD \$130.7. If, however, the hypothesized channel of impact going through an increased productivity in the agricultural production is indeed the primary channel, the effects on household welfare could easily be of a more permanent nature.

7. Conclusion

Using a randomized control trial, this paper evaluates the impact of introducing savings and loan associations in northern Malawi. Savings and loan associations (VSLAs) are designed to increase financial intermediation by mobilizing savings in the local community, which can be re-invested in the local community and increase economic activity.

We found evidence that the intervention increases household welfare: households in the poorest, most rural areas can benefit from financial services organized locally. VSLAs and other similar savings groups are common throughout rural Africa, where there are at least 60,000 such savings groups. To this date, however, very few thorough evaluations of the impact of the intervention exist.⁴¹

By randomly allocating the intervention to twenty-three among forty-six villages in the traditional authority of Mwirang’ombe in the Northern Region of Malawi, we are among the first to credibly document the effects of such an intervention. We evaluate the intervention on a set of outcomes that were predefined by the project team. In order to understand the potential mechanisms, we then investigate impacts on other intermediate outcomes.

⁴¹ We know of one other similar RCT-based study currently being conducted in Malawi by the Innovations for Poverty Action (IPA), as mentioned in the introduction.

The analysis shows that the randomization was indeed successfully implemented with savings and credit activity through VSLAs increasing dramatically in the treatment area. Importantly, the intervention affected a number of the predefined outcome variables: VSLA savings, food security, total consumption, and number of rooms in the dwelling. These results are robust to a number of different specifications, including adjusted regressions and instrumental variable estimation.

Households in treatment villages increased the number of meals they consumed the day before the survey by 0.13 ($p < 0.05$), and the value of total consumption measured using USAID's Poverty Assessment Tool for Malawi increased by 3.2% ($p < 0.05$). We do not find any effect on the length of the hungry period—that is, the period when the household ate fewer than three meals per day—nor do we find an increase in food consumption when measured using weekly recall on seventeen food items. With regard to housing, we find that the number of rooms increased by 0.15 ($p < 0.05$), but we do not find any signs of improvements in housing quality.

Apart from giving out loans, groups distributed members' savings with interest once per year. Participants reported that they spent their savings on agriculture and loans on small-scale business. After investigating overall impacts, we followed this money by looking at effects on business activity as well as agricultural inputs and outputs. We find indications that agricultural investments increases, specifically we find that the use of fertilizer and irrigation ($p < 0.1$ and $p < 0.05$) and that the value of the total maize harvest increased ($p < 0.01$). Although one of our specifications showed that households in treatment villages started slightly more businesses than households in control villages ($p < 0.1$), we find no effect on total income from these businesses. As such, agricultural investment carries more weight in explaining the results.

The findings should be of interest to practitioners, donors, and governments in developing countries alike. Apart from providing a credible assessment of the impact of the VSLA intervention in question, this analysis also offers insights into how these groups work.

The fact that agriculture seems to play a crucial role is important and somewhat surprising in this respect. The study thus adds to the literature on agricultural finance in general, and in Malawi in particular. Since 2006, the government of Malawi has subsidized agricultural inputs to smallholder farmers, but the sustainability of this program has often been called into question. The present analysis, in particular the cost-effectiveness analysis above, suggests that VSLAs can yield benefits and at a very low cost. Our results thus supplement findings from other studies aimed at the subsidizing of fertilizer, e.g. Duflo *et al.* (2011).

The above conclusions have limitations, however, which should be kept in mind. Out of the 80,000 VSLAs across the world, we have only studied a few. The results are specific to the cultural and economic context in which they appear, and a similar intervention in a change in location, another implementing organization, or just a different time period might give different results. Future assessments from other places, including the ones we know are under way, will shed light on the extent to which these findings can be generalized.

Surviving on one dollar a day is hard. But making ends meet in rural Malawi, where farming is the basis of subsistence and rain is uncertain, is even harder. Supporting these communities is difficult because they are far apart and because the threat of dependence is ever present. In that light, the positive effects we report above of small groups that do nothing but handle financial matters, are encouraging.

8. Appendix A: Estimating Food Consumption

To evaluate the poverty level of the households surveyed in the Karonga Assessment of Vulnerability (KAV), a number of different poverty indicators were included: USAID's Poverty Assessment Tool (PAT), the Progress Out of Poverty Index (PPI) developed by the Grameen Foundation, as well as a food-consumption module. The following describes the construction of the food-consumption module in the KAV and the subsequent steps undertaken to estimate the value of total food consumption. The challenges encountered in obtaining an accurate measure of consumption are identified and discussed, along with the implications for assessment of poverty through consumption-based survey techniques.

A1: How to Measure Consumption

Since cross-country comparisons of poverty require a uniform definition of poverty, measuring consumption is central to development research in order to determine who consumes below some predetermined level and can be labeled poor (Deaton and Grosh 1998). But measuring consumption accurately is not an easy task, especially when taking the cost of data collection into account. In a recent experiment, the World Bank tested the accuracy and cost effectiveness of a number of different ways of collecting information about household consumption: recall-based methods versus keeping diaries, the importance of the length of the recall period, the number of items that should be included in a survey module, and (finally) the effect of having supervisors monitoring the completion of diary entries (Beegle *et al.* 2010). The study concluded that a survey with a seven-day recall period of the consumption of a long list of specific food items provided the most accurate and cost-effective way of measuring consumption. This is also the method often applied by the World Bank in its Living Standards and Measurement Surveys (LSMS). However, a full-scale consumption module would be impossible to implement given the budgetary restrictions in the

current survey. Instead, we decided to identify the most important food items in the area, based on the most recent World Bank LSMS from Malawi—the Integrated Household Survey from 2004/05 (IHS2)—and ask about the consumption of these in a seven-day recall period in order to get a proximate measure of total consumption and thus assess the poverty level of the households.

A2: Items Included in Questionnaire

The World Bank regularly carries out surveys to assess the poverty level and poverty headcounts across developing countries. To obtain as accurate a measure of poverty as possible, an extensive consumption module is included, which typically inquires about the consumption of a comprehensive list of food and non-food items over the past week, or month(s). The seventeen items included in our questionnaire were selected based on one of these surveys: the 2004/05 Malawi Integrated Household Survey (IHS2) carried out by the National Statistics Office of Malawi for the World Bank. Using this dataset, we identified the most important items of consumption for the Karonga District and Mwirang’ombe Traditional Authority (TA), i.e., the consumption items with the highest share of total consumption. Since there were only forty observations within Mwirang’ombe TA—the specific TA in which the KAV survey was to be conducted—we decided to identify the food items using data from the entire Karonga district in which Mwirang’ombe TA is one of six TAs.

As can be seen from table A1, these seventeen items make up 89% of total food consumption in the Karonga district from the 2004/05 Malawi HIS-2. For the subsample of TA Mwirang’ombe, the seventeen items make up 91% of total food consumption, although an item such as “wild green leaves” which made up 1.8% of total food consumption for the Mwirang’ombe TA subsample was not included.

Table A1. Top Foods in Karonga District, Measured as Share of Total Food Consumption

Rank	Item #	Description	Share	Cumulative
1	503	Fresh fish	0.2614	0.2614
2	502	Dried fish	0.1726	0.4340
3	102	Fine maize flour	0.0956	0.5296
4	202	Cassava flour	0.0762	0.6058
5	101	Normal maize flour	0.0481	0.6539
6	404	Nkwani	0.0420	0.6959
7	105	Green maize	0.0273	0.7232
8	106	Rice	0.0254	0.7486
9	810	Salt	0.0221	0.7707
10	302	Bean, brown	0.0219	0.7926
11	803	Cooking oil	0.0181	0.8107
12	507	Chicken	0.0178	0.8285
13	201	Cassava tubers	0.0165	0.8450
14	405	Chinese cabbage	0.0145	0.8595
15	801	Sugar	0.0128	0.8723
16	203	White sweet potatoes	0.0096	0.8819
17	408	Tomatoes	0.0093	0.8913

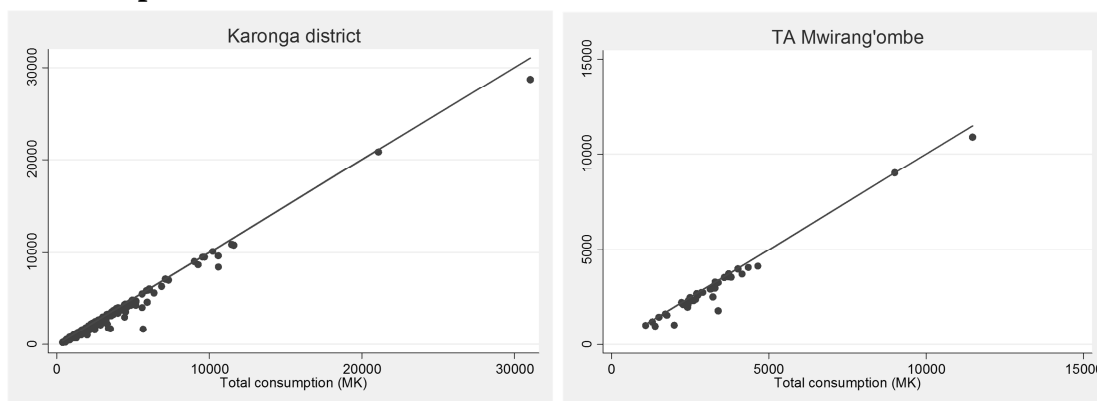
Note: Own calculations based on IHS2 data. Item # refers to the item number in the IHS2 questionnaire. The “share” column reports the share of total household consumption made up by the specific item.

The following provides some descriptive statistics on the food consumption in the Karonga district and TA Mwirang’ombe captured by the seventeen items included in the questionnaire. Within the Karonga district, the seventeen items captured on average 86% of the total food consumption, and within TA Mwirang’ombe they captured 91%. Figure A1 shows that the seventeen items captured total food consumption very well, irrespective of the total food consumption. The seventeen items should provide a good approximation for the total household consumption, for poor and less poor households alike.

Table A2. Descriptive Statistics on IHS2 Dataset

	N	Mean	Median	Min	Max
Karonga District					
Total food consumption	240	3,058	2,385	373	31,023
KAV consumption	240	2,726	2,075	208	28,690
KAV share	240	0.86	0.89	0.29	1.00
TA Mwirang'ombe					
Total food consumption	40	2,800	3,157	1,087	11,472
KAV consumption	40	2,918	2,573	977	10,885
KAV share	40	0.91	0.94	0.53	1.00

Note: Own calculations based on IHS2 data

Figure A1. Scatter Plot of Total Food Consumption and Consumption of Food Items Included in KAV

Note: Own calculations based on IHS2 dataset. Dots indicate households, and the line has a slope of 1. Pair-wise correlations of the two measures are 0.991 and 0.987 for the Karonga district and TA Mwirang'ombe respectively.

A3: Units of Measurement

Having identified the food items to include in the questionnaire, the next step was to determine the way in which food consumption should be measured. We followed the IHS2 by using the eighteen units of measurement included in the IHS2. By using these we could rely on the conversions into grams for each unit-item combination collected by the World Bank in the IHS2. One notable exception was the unit of basin/pot, which was added to the units of measurement by us and was not part of the IHS2 questionnaire. How this is dealt with in each of

the methods used is described below. Table A3 below shows the measurement units used in the 2009 data collection.

Table A3. Measurement Units Used in the 2009 Data Collection

Unit Code
1. Kilogramme
2. Gram
3. Pail (small-5littres)
4. Pail (large-25litres)
5. No. 10 plate
6. No. 12 plate
7. Bunch
8. Piece/cob
9. Heap
10. Bale
11. Basket (<i>Dengu</i>) shelled
12. Basket (<i>Dengu</i>) unshelled
13. Liter
14. Milliliter
15. Basin/pot
16. Cup
17. Spoon
18. Tin
19. Other (specify)

However, due to problems in these conversions (as will be described below), we used a different approach in the 2011 data collection. There we used two types of measurement units: metric units directly, and pictures of units for which we, ourselves, measured the conversions into kilogram. The pictures are shown at the end of this appendix.

The use of pictures ensures a common understanding of the terminology used between the researcher, the interviewer, and the respondent. Pictures were taken as part of the piloting prior to the 2011 data collection at markets within the local area and were the units commonly used when selling and buying goods in this particular context. From the collection of these pictures and their corresponding

conversions into kilograms, it became evident that often very few units of measurement were used, but each food item had a particular unit in which it was usually measured when traded.

Table A4. Metric Units Used in the 2011 Data Collection

<u>Unit Code</u>
1. Kilogram
2. Gram
3. Liter
4. Milliliter

A4: Calculating Value of Consumption in 2009

In IHS2, the value of total consumption was calculated in the following way. The actual expenditure was used for households who purchased the particular food item. Households that consumed items produced at home or which they received as a gift had these values imputed using the median price in the enumeration area. This median price was calculated using the data from households that purchased the specific item. If less than seven households purchased the item in the enumeration area at the specific time of the interview, either time or area was expanded before evaluating the median price (World Bank 2006).

We initially planned to estimate the value of consumption in a similar fashion, but identified a number of serious shortcomings in the conversion from the nineteen units of measurement into kilograms. These shortcomings, and our solutions to them, are described below. In particular, we calculate the value of food consumption by using specific unit-item prices collected through our survey.

Problems in Using IHS2 Conversions

There are three key problems with the calculation of total consumption in IHS2. Available from the IHS2 is a conversion dataset, where the weight in grams of a specific item measured in a specific unit is available. This conversion dataset is not complete, however, and does not contain information about every single

item/unit combination. For instance, a conversion for normal maize flour is only available for pail (small and large), no ten and twelve plates, piece/cob, baskets (shelled and unshelled), cup and tin. This incompleteness results in 359 observations about weekly consumption of particular items for which we have no knowledge of the conversion into kilograms (out of a total of 17,765). Almost half of these (157) are due to a missing conversion for a tin of rice. On top of these 359 observations, the consumption in an additional 167 observations was stated using the “other” category, for which we also have no means of converting the consumption into kilograms. Finally, the basin/pot category was used in 2,250 observations. In total, we lack conversions for 15.6% of the observations, the majority of these being due to the inclusion of the basin/pot measurement unit, which was not part of the IHS2.

When calculating the value of consumption, it became evident that there are substantial problems even among the item/unit combinations for which the IHS2 contained information. Most importantly, a “tin” of maize contained 33 grams according to the IHS2 conversions. In the area where we worked, villagers told us that a 50-kilogram bag of maize typically contained 3 tins of maize—thus a conversion of approximately 17 kilograms. When looking at the prices paid in IHS2, the median price for a kilogram was MK 38 and the corresponding price for a tin of maize was MK 600. This suggests that conversion of 16 kilograms for a tin of maize comes very close to the anecdotal evidence. We conclude that the conversion used in the IHS2 for a tin of maize, which is the main staple crop in Malawi, is most likely wrong to an extent that makes it unusable.

Similar problems were present in the conversion of a piece of fish—both dried and fresh. The conversion suggested that a piece of dried fish should weigh 965 (865) grams. But where the median price stated for a kilogram of dried fish was MK 63 (120), a piece of fresh fish cost MK 10 (13). Anecdotally, the fish we saw at the markets in the area were typically very small (see also the fish in the picture below). Furthermore, from conversations with the local fishermen, we learned that the Chambo—the most typical and desired eating fish

from Lake Malawi (typically about 30 centimeters long and weighing approximately 1 kilogram)—was no longer caught in the area, only in the southern parts of the lake (approximately 200 kilometers from the survey area).

Since reporting the consumption of maize flour in tins and the number of fresh and dried fish (the piece unit) is quite common in our survey, these three problems alone had significant impacts on the estimated value of total consumption. Table A6 below shows the bias in the consumption of these items compared to the actual methodology we applied, which will be described below.

Using a Different Approach

An alternative to relying on the IHS2 conversion is to use the conversions implicit in the collected data. For example, some respondents stated the price paid for the quantity of maize bought in tins, some stated the bought quantity in kilograms, and some stated the quantity bought in pails. Thus, from our data we can generate the median item/unit specific prices and evaluate the household consumption by these prices. Due to the relatively large number of item/unit prices to calculate, these are done across the entire survey area.

We therefore estimate the value of aggregate consumption by valuing the reported consumption by item/unit-specific median prices from the dataset. This method has its own issues, two of which are noteworthy. For one, we do not have price information on all item/unit combinations. This results in missing prices for 283 observations. Secondly, and perhaps more importantly, since there is a large number of item/unit-specific prices to estimate and we only have information about actual expenditures from the long questionnaire data, the item/unit-specific median prices are often based on a very low number of observations. Table A5 below shows the number of observations on which the estimated price for each item/unit combination is estimated. The colors indicate whether the price is based on greater or fewer than seven observations. This is the minimum number of observations required by the World Bank in their estimation of area-specific prices.

Approximately half of the estimated prices—77 of 156—are based on fewer than seven observations. However, these item/unit combinations are also more rarely used in reporting the actual consumption.

Table A5. Table on Number of Observations on which Item/Unit Prices Are Based

	Unit																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Item	1	2		7	13	1							1		2			19	
	2	3		6	8								4		2			21	
	3		1					40		1			1		3	1			
	4	64	1	30	9	9	20		1	1					51	11		10	
	5	1		2		1		18	36	1	1	1			22			1	
	6	10		16	4	1				1		1			29	1		11	
	7	5		9	4	1	3	1	3						52	1		4	
	8																		
	9	105	3	1	1	67	77		8	4		7			32	70		1	2
	10	0					140		12						10				
	11	6				1	159	28	12						14			1	17
	12	10		32	1	23	24	2	20	317	1				118	1		1	
	13	8		22	2	9	17	1	87	241		1			47				2
	14	9	2	4		10	16	1	57	286					31				2
	15	15						7	3										7
	16	625	23	1					1	4				1	1	9			10
	17	11	2	1	1	1							240	247	1	23	4		30
	18	98	148	1	1	16	11	1		8	2			1	1	469	7		26

For the item/unit combinations for which we have no price information, the value of consumption is estimated using the median consumption per adult equivalent of the specific item.

Table A6 shows the estimated food consumption per adult equivalent per week using the preferred methodology. The final columns show the same descriptive statistics for the estimated consumption relying on the conversion tables provided by the World Bank. It is evident that the value of consumed normal and fine maize flour is especially implausible when using the World Bank conversions.

Table A6. Summary Statistics on Value of Consumption per Adult Equivalent, 2009 Data

Item	N	Item/Unit-Specific Prices				IHS2 Conversions			
		Mean	Median	Min	Max	Mean	Median	Min	Max
Normal flour	1,752	38	0	0	583	311	0	0	58,997
Fine flour	1,752	88	81	0	1,256	3,261	86	0	283,050
Green maize	1,752	6	0	0	388	6	0	0	388
Rice	1,752	57	35	0	887	56	27	0	1,442
Cassava tubers	1,752	22	9	0	1,258	23	9	0	697
Cassava flour	1,752	33	16	0	552	43	3	0	1,726
Sweet potatoes	1,752	18	0	0	1,188	16	0	0	1,168
Brown beans	1,752	19	0	0	735	22	2	0	444
Pumpkin leaves	1,752	5	0	0	256	8	0	0	454
Chinese cabbage	1,752	6	0	0	1,282	12	0	0	1,282
Tomatoes	1,752	30	23	0	1,500	30	22	0	2,000
Fresh fish	1,752	42	26	0	701	51	24	0	3,559
Dried fish	1,752	26	18	0	333	28	17	0	350
Chicken	1,752	63	0	0	1,351	52	0	0	955
Sugar	1,752	31	28	0	341	32	29	0	341
Cooking oil	1,752	28	20	0	462	32	19	0	442
Salt	1,752	11	8	0	153	12	8	0	249
Total consumption	1,752	523	442	0	3,350	3,995	696	0	284,226

A5: Calculating Value of Consumption in 2011

Following the problems encountered using the units and conversions based on the IHS2 methodology in the 2009 data collection, the consumption module was changed for the 2011 survey. As described in the previous section on units of measurement, the 2011 survey included fewer unit options, but combined these with area-specific pictures of the most commonly used units of measurement for each of the food items.

In estimating the value of consumption in the 2011 survey, the following steps were taken: 1) Converting all purchased quantities into kilograms using the pictures and our own collected unit conversions; 2) Estimating the median price per kilogram in the village when at least seven observations are available. If fewer than

seven observations were available, we expanded the area to include all villages under the same group village headman (GVH) or finally included all observations in the survey.⁴² And 3) Estimating the value of consumption based on the median prices calculated.

Table A7 shows the descriptive statistics on the consumption per adult equivalent of each of the items. The descriptive statistics further validate the use of item/unit-specific prices in the 2009 survey. The descriptive statistics of the estimated consumption pattern using the picture codes in 2011 rather closely resembles that of the 2009 survey when using the preferred methodology.

Table A7. Summary Statistics on Value of Consumption per Adult Equivalent, 2011 data

Item	N	Mean	Median	Min	Max
Normal maize flour	1,735	12	0	0	348
Fine maize flour	1,735	118	105	0	997
Green maize	1,735	5	0	0	413
Rice	1,735	98	65	0	1,240
Cassava tubers	1,735	16	0	0	476
Cassava flour	1,735	44	11	0	1,512
Sweet potatoes	1,735	37	17	0	1,284
Brown beans	1,735	13	0	0	301
Pumpkin leaves	1,735	17	8	0	1,505
Chinese cabbage	1,735	6	0	0	257
Tomatoes	1,735	43	28	0	1,553
Fresh fish	1,735	51	27	0	1,002
Dried fish	1,735	59	21	0	3,981
Chicken	1,735	65	0	0	1,111
Sugar	1,735	47	41	0	848
Cooking oil	1,735	33	22	0	850
Salt	1,735	10	8	0	93
Total HH consumption	1,735	691	562	58	4,846

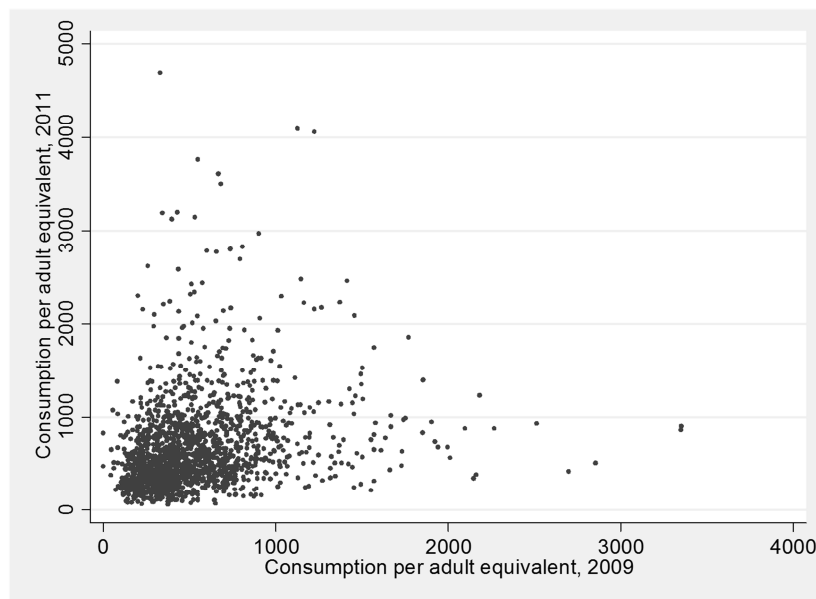
⁴² This is similar to the methodology applied by the World Bank when estimating area-specific prices. We only have price information on the 834 households that were administered the long questionnaire. The short questionnaire only asks for the quantity consumed.

A6: Comparing Consumption Across Time

Given the differences in the methodology used to measure the value of consumption in 2009 and 2011, it is of interest how the two measures compare. Although we are unable to compare the two measures directly, we can see whether there is any correlation over time between the two rounds. There are, of course, a number of factors affecting the household consumption pattern over time (one of these potentially being the VSLA intervention in the treatment villages), but there should nevertheless be no systematic differences between the two measures.

The following figure therefore plots the estimated consumption per adult equivalent per week in 2009 against the same measure from 2011. The figure shows a far-from-perfect correlation between the two measures.⁴³ And perhaps more importantly, there seems to be a lot of noise in the measures. The difference between the 2009 and 2011 food consumption levels, and the large variance, could simply reflect the general problems with obtaining an accurate measure of consumption through survey methods, as mentioned initially.

Figure A2. Estimated Value of Consumption (MK) per Adult Equivalent, 2009 and 2011



⁴³ The correlation coefficient is 0.239.

A7: Converting Value of Consumption into U.S. Dollars

Estimating the value of consumption is the first step in evaluating the level of poverty. In order to compare these figures across countries, it is necessary to convert the local currency into a common currency. But the law of one price required for nominal exchange rates to ensure equal prices across countries is seldom justified, as the nominal exchange rates are also affected by a various factors (the level of political stability, for instance). Using the nominal exchange rate to make this conversion would thus distort a cross-country comparison of incomes or consumption. This was not the main focus of this survey, but nevertheless the discussion is relevant for comparing the poverty level of the survey area with other areas.

Instead of comparison based on nominal exchange rates, purchasing power parity (PPP) exchange rates are often applied. These convert any currency into a hypothetical international dollar having the same purchasing power as a U.S. dollar did in the USA in 2005. In doing that, standard PPP uses a generic bundle of goods. One problem with this is that the bundle of goods used to create the PPP exchange rates does not necessarily reflect the bundle of goods usually consumed by the poor households. If the price of the goods usually consumed by poor households differs systematically from the prices of the goods included in the common bundle of goods compared across countries, this provides a distorted image of the consumption or income patterns of the poor households. To overcome this, Deaton and Dupriez (2011) construct an alternative set of poverty-adjusted PPP exchange rates using bundles of goods and prices from developing-country household surveys. For this reason, the measure is only available in the countries that have poor households in the global sense—that is, households consuming less than \$1.25 a day. Importantly, this excludes the USA, which makes a poverty-adjusted PPP conversion into 2005 US dollars impossible. However, in the words of the authors:

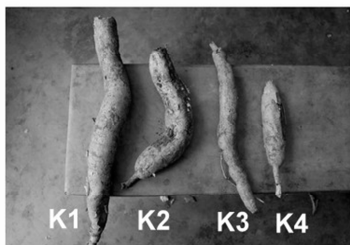
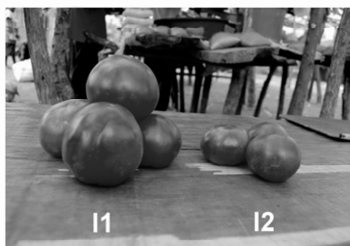
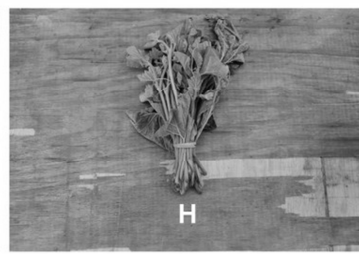
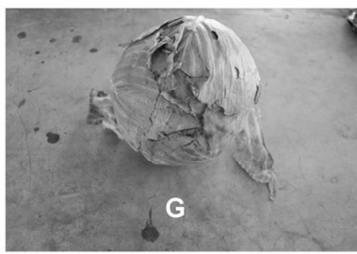
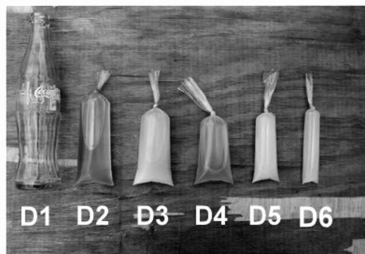
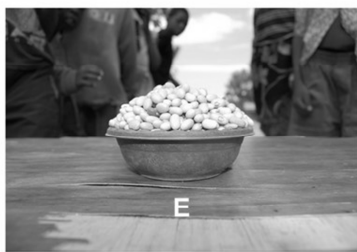
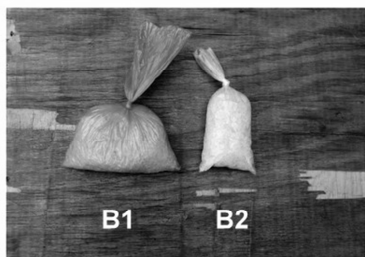
“While we recognize that it is inevitable that people will want such numbers, a good reason for not calculating them is that the structure of the United States, or of other advanced economies, is quite different

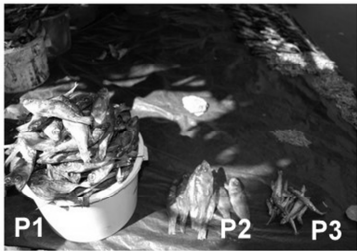
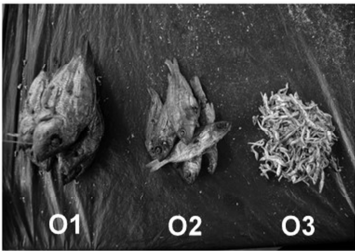
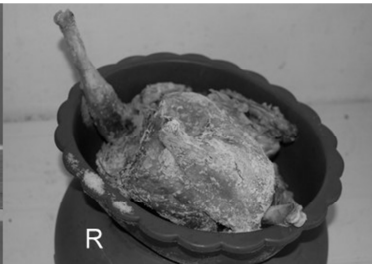
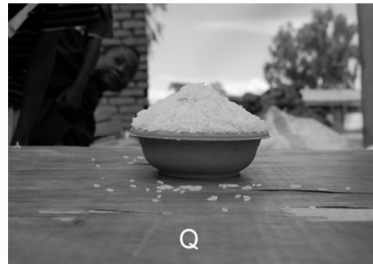
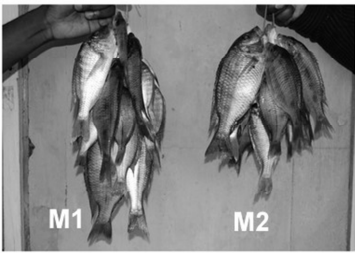
from the structures of the economies where the global poor live, so that index numbers that compare the two are subject to a great deal of uncertainty and vary greatly.” (Deaton and Dupriez 2011:159)

They do, however, provide a figure for the conversion into 2005 USD, which is the one used here. Table A8 below compares the nominal exchange rate with the PPP and the poverty-adjusted PPP exchange rates for Malawi in the period from 2009 to 2011.

Table A8. Exchange Rates for Malawi, 2005-2011

	2005	2006	2007	2008	2009	2010	2011
Nominal Exchange Rate	116.84	135.54	139.72	140.91	141.75	150.73	156.94
PPP	56.92	62.85	65.97	69.06	75.14	79.41	82.85
4P	62.88	71.66	77.36	84.10	91.19	97.94	105.41





9. References

- Allen, H. and Panetta, D. 2010. *Savings Groups: What Are They?* Washington DC: The SEEP Network.
- Allen, H. and Staehle, M. 2007. *Programme Guide - Field Operations Manual version 3.1*. Stuttgart: VSL Associates, available from <http://vsla.net> [Accessed October 2012].
- Anderson, S. and Baland, J.-M. 2002. The Economics of ROSCAs and Intrahousehold Resource Allocation. *Quarterly Journal of Economics*, 117(3), pp. 963-995.
- Angrist, J.A. and Imbens, G.W. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), pp. 444-455.
- Angrist, J.D. and Pischke, J.-S. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.
- Aryeetey, E. 2005. Informal Finance for Private Sector Development in Sub-Saharan Africa. *Journal of Microfinance*, 7(1), pp. 14-38.
- Aryeetey, E. and Steel, W. 1994. Informal savings collectors in Ghana: can they intermediate? *Finance and Development*, 31(1), pp. 36-41.
- Ashe, J. 2002. A Symposium on Savings-Led Microfinance and the Rural Poor. *Journal of Microfinance*, 4(2), pp. 127-135.
- Ashraf, N., Karlan, D. and Yin, W. 2006. Tying Odysseus to The Mast: Evidence From a Commitment Savings Product in the Philippines. *Quarterly Journal of Economics*, 121(2), pp. 635-672.
- Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E. and Harmgart, H. 2011. Group lending or individual lending: Evidence from a randomized field experiment in Mongolia. *EBRD Working Paper*. European Bank for Reconstruction and Development, 48.

- Banerjee, A. and Duflo, E. 2009. The Experimental Approach to Development Economics. *Annu. Rev. Econ.*, 1(1), pp. 151-178.
- Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. 2010. *The miracle of microfinance? Evidence from a randomized evaluation*. Working Paper: Financial Access Initiative and Innovations for Poverty Action.
- Bloom, H.S. 1995. Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation review*, 19(5), pp. 547-556.
- Bouman, F. 1995. Rotating and accumulating savings and credit associations: A development perspective. *World Development*, 23(3), pp. 371-384.
- Bruhn, M. and McKenzie, D. 2009. In pursuit of balance: randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4), pp. 200-232.
- Carter, M.R. and Zimmerman, F.J. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics*, 71(2), pp. 233-260.
- Casey, K., Glennerster, R. and Miguel, E. 2012. Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. *The Quarterly Journal of Economics*, 127(4), pp. 1755-1812.
- Chattopadhyay, R. and Duflo, E. 2004. Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica*, 72(5), pp. 1409-1443.
- Collins, D., Morduch, J., Rutherford, S. and Ruthven, O. 2009. *Portfolios of the Poor: How the World's Poor Live on \$2 a Day*. Princeton, NJ: Princeton University Press.
- Conning, J. and Udry, C. 2007. Rural financial markets in developing countries. In P. Pingali and R. Evenson (eds.) *Handbook of Agricultural Economics*. The Netherlands: Elsevier, pp. 2857-2908.

- Copestake, J., Duvendack, M., Hooper, L., Loke, Y., Palmer-Jones, R. and Rao, N. 2011. What is the evidence of the impact of microfinance on the well-being of poor people? London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
- Dale, R. 2003. The logical framework: an easy escape, a straitjacket, or a useful planning tool? *Development in practice*, 13(1), pp. 57-70.
- Daley-Harris, S. 2009. *State of the Microcredit Summit Campaign Report 2009*. Washington DC: Microcredit Summit Campaign.
- Deaton, A. 1991. Saving and Liquidity Constraints. *Econometrica: Journal of the Econometric Society*, 59(5), pp. 1221-1248.
- Deaton, A. 2001. Counting the world's poor: problems and possible solutions. *The World Bank Research Observer*, 16(2), pp. 125-147.
- Deaton, A. and Dupriez, O. 2011. Purchasing power parity exchange rates for the global poor. *American Economic Journal: Applied Economics*, 3(2), pp. 137-166.
- Deaton, A. and Grosh, M. 1998. The consumption module in the living standards measurement survey. In P. Glewwe and M. Grosh (eds.) *Designing Household Survey Questionnaires for Developing Countries*. Washington DC: World Bank.
- DeLong, J.B. and Lang, K. 1992. Are all economic hypotheses false? *The Journal of Political Economy*, 100(6), pp. 1257-1272.
- Demirgüç-Kunt, A. and Klapper, L. 2012. *Measuring Financial Inclusion: The Global Findex Database*. World Bank Policy Research Working Paper No. 6025. Washington, D.C.: World Bank, Available at SSRN 2043012.
- Diamond, D.W. 1984. Financial intermediation and delegated monitoring. *The Review of Economic Studies*, 51(3), pp. 393-414.
- Duflo, E., Glennerster, R. and Kremer, M. 2007. Using randomization in development economics research: A toolkit. In T.P. Schultz

- and J.A. Strauss (eds.) *Handbook of development economics*. Elsevier, pp. 3895-3962.
- Duflo, E., Kremer, M. and Robinson, J. 2011. Nudging farmers to use fertilizer: theory and experimental evidence from Kenya. *American Economic Review*, 101(6), pp. 2350-2390.
- Freedman, D.A. 2008a. On regression adjustments in experiments with several treatments. *The annals of applied statistics*, 2(1), pp. 176-196.
- Freedman, D.A. 2008b. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2), pp. 180-193.
- Freixas, X. and Rochet, J.C. 2008. *Microeconomics of Banking*. Cambridge, Mass.: MIT Press.
- Fudenberg, D. and Levine, D.K. 2006. A Dual-Self Model of Impulse Control. *American Economic Review*, 96(5), pp. 1449-1476.
- Gelman, A., Hill, J. and Yajima, M. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of research on educational effectiveness*, 5(2), pp. 189-211.
- Gerber, A.S. and Green, D.P. 2012. *Field experiments : design, analysis, and interpretation*, 1. ed. New York: W. W. Norton.
- Gleerup, M., Larsen, A., Leth-Petersen, S. and Togeby, M. 2010. The Effect of Feedback by Text Message(SMS) and Email on Household Electricity Consumption: Experimental Evidence. *Energy Journal*, 31(3), pp. 113-132.
- Glewwe, P. and Jacoby, H. 2000. Recommendations for Collecting Panel Data. In M. Grosh and P. Glewwe (eds.) *Designing household survey questionnaires for developing countries : lessons from 15 years of the Living Standards Measurement Study*. Washington DC: World Bank, pp. 275-314.
- IRIS Center 2012. Poverty Assessment Tool Accuracy Submission USAID/IRIS Tool for Malawi. Maryland: University of Maryland.

- Kennedy, P.E. 1981. Estimation with correctly interpreted dummy variables in semilogarithmic equations. *American Economic Review*, 71(4).
- Lechner, M. 2008. A Note on the Common Support Problem in Applied Evaluation Studies. *Annals of Economics and Statistics / Annales d'Économie et de Statistique*, (91/92), pp. 217-235.
- Leth-Petersen, S. 2010. Intertemporal consumption and credit constraints: Does total expenditure respond to an exogenous shock to credit? *The American Economic Review*, 100(3), pp. 1080-1103.
- Lin, W. 2013. Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *Annals of applied statistics*, 7(1), pp. 295-318.
- McKenzie, D. 2012. Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2), pp. 210-221.
- O'Donoghue, T. and Rabin, M. 1999. Doing it now or later. *American Economic Review*, 89(1), pp. 103-124.
- Parsons, V. 2005. Stratified sampling. In N.C.F.H. Statistics (ed) *Encyclopedia of biostatistics*. Hyattsville: John Wiley and Sons, Ltd.
- Pfeffermann, D. 1993. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 61(2), pp. 317-337.
- Rasmussen, O.D. 2012. Small groups, large profits: Calculating interest rates in community-managed microfinance. *Enterprise Development and Microfinance*, 23(4), pp. 298-318.
- Rasmussen, O.D., Malchow-Møller, N. and Andersen, T.B. 2011. Walking the talk: the need for a trial registry for development interventions. *Journal of Development Effectiveness*, 3(4), pp. 502-519.

- Reynolds, C.L. and DesJardins, S.L. 2009. The use of matching methods in higher education research: Answering whether attendance at a 2-year institution results in differences in educational attainment. In J.C. Smart (ed) *Higher education: Handbook of Theory and Research*. The Netherlands: Springer, pp. 47-104.
- Rosenbaum, P.R. and Rubin, D.B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, pp. 41-55.
- Rutherford, S. 2001. *The poor and their money*. Oxford: Oxford university press.
- Rutherford, S. 2009. *The Poor and Their Money*, 2nd ed. London: Practical Action Publishing.
- Smith, J. and Todd, P. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1), pp. 305-353.
- Stewart, R., van Rooyen, C., Dickson, K., Majoro, M. and de Wet, T. 2010. *What is the impact of microfinance on poor people? A systematic review of evidence from sub-Saharan Africa*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
- Thaler, R.H. 1990. Anomalies: Saving, fungibility, and mental accounts. *The Journal of Economic Perspectives*, 4(1), pp. 193-205.
- Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Boston, MA: The MIT Press.
- World Bank 2004. *Note on Construction of Expenditure Aggregate and Poverty Lines for IHS2*. Washington, D.C.: World Bank, available from <http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1181743055198/3877319-1181928149600/ihs2agg.pdf> [Accessed May 2013].
- Zimmerman, F. and Carter, M. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and

subsistence constraints. *Journal of Development Economics*, 71(2), pp. 233-260.

Chapter 2

Can Microfinance Reach the Poorest: Evidence from a Community-Managed Microfinance Intervention

Can Microfinance Reach the Poorest: Evidence from a Community-Managed Microfinance Intervention

Jonas Helth Lønborg
University of Southern Denmark

Ole Dahl Rasmussen
University of Southern Denmark and DanChurchAid

Abstract: Reaching the poorest is an important objective in many development interventions, and microfinance is no exception. We review performance indicators for effectiveness of targeting described in the literature and suggest a new metric in order to account for extent and severity of poverty as well as the income distribution among the poor. When applying this to a panel dataset from a community-managed microfinance intervention in Northern Malawi, we find regressive targeting: Participants are less poor than the general population in the area. This result is robust to several ways of measuring poverty. Additionally, we decompose the participation process into six steps and provide evidence that the awareness campaign never reaches a large share of the population.

Keywords: targeting, microfinance, Africa, VSLA

Acknowledgements: We thank IKI for collecting the data, participants in the seminar at the Rockwool Foundation Research Unit, Nikolaj Malchow-Møller and Thomas Barnebeck Andersen for comments on earlier drafts, and DanChurchAid, the Rockwool Foundation Research Unit, and the Danish Agency for Science, Technology and Innovation for financial support. Any viewpoints and errors are our own

Introduction

Together with economic growth, poverty reduction is perhaps the most agreed-upon goal for development aid. Targeting interventions and resources toward the poorest is often viewed as a necessary means to reaching this goal. At the same time, however, an increasing number of development interventions require strong involvement from and capacity of participants. It is likely that this limits the outreach of the interventions. To guide the implementation and design of interventions we need more knowledge about participation in interventions which require a high degree of involvement by participants.

Microfinance, which includes the provision of loans, savings, and insurance, is a case in point. Targeting is widely used in microfinance as a means to deepen outreach by, for example, the Grameen Bank and BRAC (Bandiera *et al.* 2011), and there has been increasing focus on avoiding "mission drift", whereby programs include richer people (Christen 2001, Cull *et al.* 2007, Hermes *et al.* 2011). At the same time, however, a common opinion is that microfinance does not work for the poorest, a finding which is confirmed by early studies (Hulme 2000, Navajas *et al.* 2000). Microcredit, in particular, requires a high degree of involvement as well as prior skills from the loan takers. Borrowers need to be able to use a loan to create income, keep track of repayment schedules, and possess basic financial literacy.

In this paper we investigate to what extent it is possible to reach the poorest with a program that requires a high degree of involvement by participants. We analyze participation in the context of one of the most poverty-focused microfinance methodologies, namely community-managed microfinance, in a rural area in one of the world's poorest countries, Malawi.

To do this, we use a panel dataset from a household survey from 885 households in northern Malawi. The first round of data was collected just before the introduction of a large-scale community-managed microfinance project. Two years later, the households were

revisited to gather information about their participation in the program.

We contribute to the literature in three ways: In addition to reporting the standard metrics of targeting effectiveness, we develop our own metric, inspired by the squared poverty gap commonly used in poverty measurement. Unlike typical targeting metrics, the new measure is sensitive to changes in the depth of poverty as well as the income distribution among the poor. Furthermore, we analyze participation in a sequential framework, assessing the necessary steps preceding participation, including project awareness and interest in participation. We picture these steps as a leaking pipeline and examine where the pipeline is leaking. This framework has been used in developed countries but only sporadically in developing countries. Finally, we illustrate the use of the new metric and the sequential approach by examining targeting in community-managed microfinance, or more specifically savings groups, a type of microfinance intervention that is highly standardized and widely used (Allen and Panetta 2010).

We find that targeting is regressive: participants are less poor than the overall population in the area. This result is even stronger when we use our own metric based on the squared poverty gap and appears in three out of four ways of measuring consumption. The exception is when we measure consumption directly using recall questions on 17 items, in which case the results point in the opposite direction but are statistically insignificant. This might be explained by measurement error on this particular variable. Asked about the reasons for not joining, non-participants report that the problem is lack of cash to fulfill the compulsory savings requirements.

The analysis of the pipeline of participation shows that the awareness campaign initially attracts both the poor and the non-poor, but that the poor are first-movers in the sense that they are more likely to join, given they received the information about the upcoming intervention.. Only later, richer households join and do so in larger numbers. In other words: the awareness campaign seems to attract a different group of people than those who end up joining.

The rest of the article is structured as follows. The following section describes the intervention implemented in the area, the so-called Villages Savings and Loans Associations (VSLAs). Section three provides an overview of the existing targeting literature. Section four reviews targeting measures used in the literature and develops a new metric based on the squared poverty gap. Section five explains our sequential approach. In sections six and seven we present the data and our empirical strategy before turning to the results in section eight. The final section concludes and provides policy recommendations based on the results.

The Intervention

The microfinance intervention in our study is a community-managed microfinance program called village savings and loan associations (VSLAs). VSLAs are a form of accumulating savings and credit association (following the definitions used by e.g. Bouman 1995), where villagers meet every week and contribute an amount to a common pool of funds. The procedures of setting up and running these groups are thoroughly documented in a set of manuals (Allen and Staehle 2007). Key characteristics are that no external funds are provided, so all loans are made using participants' savings. There are lower and upper limits to the amount that it is possible to save at each meeting. Credit is provided to members at an interest rate set by the group, typically five to ten percent per month with a three month repayment period. An association also includes a welfare fund financed by very small weekly payments by each member. The welfare fund can be invoked on certain occasions, for example the death of a family member, crop failure, or weddings. The formation of the groups is essential for the present study. In our case, the formation of VSLAs was facilitated by a local organization called SOLDEV. The implementing organization approaches the village leaders to get their approval of the project. The village leaders are asked to gather all villagers who might be interested in joining such a group at a designated time for an awareness meeting. These awareness meetings are held in the villages to inform people about the initiative. People

are asked to form groups with other villagers they trust. Subsequently training sessions are conducted by the implementing organization. During the first three months, a field officer participates in every group meeting and trains the groups in various aspects of the methodology: Electing a management committee, administering savings, giving out loans, etc. After the first three months, the group is still supervised by the field officer, although at a lower frequency. After twelve months, the groups “mature” and are no longer supervised by the implementing organization.

Targeting and Outreach: A Literature Review

Targeting is when an intervention aims to include only a specific subgroup of the population. Most of the literature on targeting rests on the assumption that interventions must reach the poorest in order to benefit the poorest (Amin *et al.* 2003, Coady *et al.* 2004). The opposing view, i.e. that increased poverty reduction does not follow from better targeting, has also been voiced. In this section we summarize these arguments before we turn to the approaches used when analyzing targeting.

One reason why targeting can lead to less poverty reduction is that it is costly to the implementer and can include hidden costs to participants in terms of conditions for participation or stigma (Ravallion 2009). Moreover, targeting methods might be inefficient. Niehaus *et al.* (2013) document how targeting on a large number of indicators in a proxy means test might improve statistical accuracy but at the same time decrease enforceability if implementers are corruptible.

Even if the ultimate focus of an intervention is poverty reduction, targeting may not be required. If the goal is poverty reduction through growth, then any intervention must focus on stimulating the economy as such. This might benefit the poorest through trickle down and does not depend on active participation by the poorest. However, whereas there is little doubt that GDI growth reduces poverty on average, there is a large heterogeneity in the existing evidence. The connection between growth and poverty reduction is uneven, and the link is

sometimes weak (Ravallion 2001). Even if economic growth is the end goal, there is still a reason to care about targeting as extreme poverty might in itself have adverse effects on growth (Ravallion 2012).

For these reasons, we believe that there is a strong case for investigating targeting effectiveness.

Measuring Targeting Effectiveness

One part of the literature on targeting evaluates different practical methods like proxy means tests, geographical targeting, and community-based targeting (Conning and Kevane 2002, Houssou and Zeller 2011, Alatas *et al.* 2012, Lang *et al.* 2013). Another part of the literature focuses on targeting effectiveness and in particular on whether interventions are successful in reaching the subgroup in question. This division of the literature is illustrated in figure 1. VSLAs are designed to reach the poorest but do not use explicit targeting methods like the ones just mentioned. For this reason, we focus the discussion on targeting effectiveness: How successful was the program in reaching the poorest? For an overview of the literature mentioned below, see table 1.

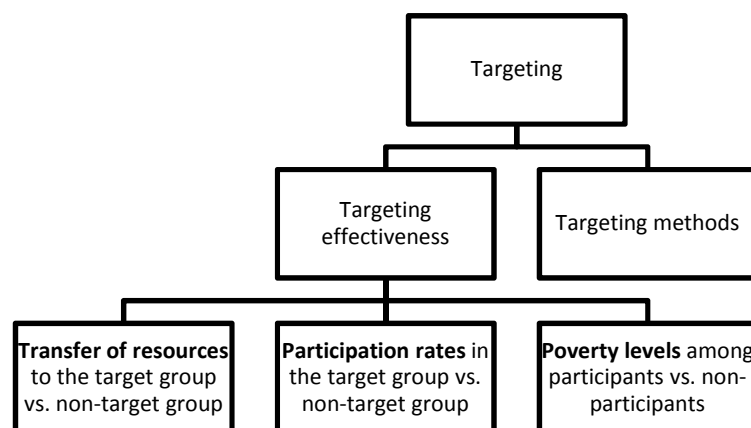


Figure 1. Sub-divisions in the literature on targeting

The literature on targeting effectiveness can be further divided into three categories according to its focus of attention. One type is concerned with the amount of resources transferred to the target

group, e.g. the poor, compared to the resources transferred to people outside the target group. A second type looks at participation rates in the two groups. A third part of the literature compares poverty levels among participants and non-participants. This is also labeled outreach analysis and is the approach we use below. Since we draw on lessons from all three approaches, we discuss key contributions from each in the following description before turning to the poverty metrics used in comparing participants to the general population in the area.

Grosh (1994) and Coady *et al.* (2004) are central to the first strand of the literature, which measures targeting by the resources transferred to the poor. Both studies compare multiple interventions across countries. To facilitate comparison they develop a generalized targeting performance indicator, which we utilize and develop further in the analysis below. Conceptually, they compare the targeting of an intervention to the common reference of neutral targeting where all subgroups of a population receive the same share of the total transfers, irrespective of the income level of the subgroup. The indicator is calculated as the share of funds transferred to the target group, e.g. the poor, divided by the target group's proportion of the overall population. A performance indicator above one indicates progressive targeting – i.e. that the poor are given preferential treatment – whereas an indicator below one indicates regressive targeting. Coady *et al.* (2004) construct a database of 122 targeted anti-poverty programs and find that a quarter of the programs exhibit regressive targeting despite ambitions of the opposite. Grosh (1994) finds that the twenty-three programs in Latin America she has information on exhibit progressive targeting.

Even though there is no transfer of resources in microfinance in general, nor specifically in community-managed microfinance, we can easily adopt the same approach when developing targeting performance indicators. Instead of using the proportion of the amount transferred, we use the average of a number of different poverty metrics for the participants divided by the same metrics of the general population in the area.

The second part of the literature compares participation rates in various ways. This is particularly useful when interventions have clear targeting criteria, e.g. the Zambian maize subsidy given to everyone with an income below K20,500 or the Jamaican food subsidy program for pregnant women, as mentioned by Cornia and Stewart (1993). The authors use these clear targeting criteria to quantify mistargeting by dividing errors into F-mistakes, which is failure to reach the entire target group, and E-mistakes, which is excessive targeting or inclusion of people from outside the target group in an intervention. Investigating primarily food subsidy schemes, Cornia and Stewart (1993) find that a targeting mechanism designed to minimize E-mistakes often increases F-mistakes at the same time. Other studies that investigate participation rates include Ravallion (2009), Handa *et al.* (2012), and Houssou and Zeller (2011). Coady *et al.* (2004), mentioned above, also use participation rates as the basis of their performance indicator whenever they cannot find information on transfers.

There are two lessons to be learned from this literature. First, participation rates are primarily useful when programs operate with clear targeting criteria. That is not the case for a program like ours, which targets the poor in general. Second, when the target group is the poor, participation rates treat all poor equally, thereby ignoring the severity of poverty. This issue is discussed further in the section on the outreach ratio.

The literature that analyzes targeting by comparing poverty levels among participants and non-participants includes Mohammed *et al.* (1999) and Amin *et al.* (2003). Both studies analyze microfinance in Bangladesh and find that participants are poorer than non-participants on average. Also, Navajas *et al.* (2000) compare poverty levels of participants in microfinance in Bolivia and find that they are just below the national poverty lines but do not belong to the very poor. The advantage in comparing poverty levels is that it allows for flexibility in the definition of poverty apart from the dichotomy poor/non-poor. We use several specific metrics from this literature, as discussed in depth below.

One way of looking at outreach in microfinance in particular is to use loan size as a proxy for the poverty level of clients (Cull *et al.* 2007, Hartarska and Nadolnyak 2007, Hermes *et al.* 2011). This approach is common despite the fact that loan size is likely to differ across sectors, possibly creating a systematic correlation since the provision of large loans may also occur to very poor people. A case in point is agricultural loans, which are likely to be larger than average, while clients might very well be poorer.

A separate issue in the literature is endogeneity since households are often surveyed after an intervention. Comparing non-participants and participants at this stage mixes pre-program differences with any positive or negative effects of participation on either levels or variances. Mohammed *et al.* (1999) and Navajas *et al.* (2000) are examples of this. Ravallion (2009) also uses post-intervention figures, but argues that pre-intervention income is equal to post-intervention income plus transfers. In the overview by Coady *et al.* (2004) it is not clear whether data is pre- or post-intervention. Our analysis avoids endogeneity since we use poverty measures collected in a survey before roll-out of the program.

The Outreach Ratio

The previous section discussed the literature on targeting in general. In this section we review specific metrics used in assessing targeting effectiveness, specifically poverty levels of participants. Our starting point is a measure of targeting effectiveness first introduced by (Coady *et al.* 2004), which we call the outreach ratio. The outreach ratio compares the actual targeting in a program with neutral targeting, i.e. a situation where the intervention reaches a representative group of the population. The advantage is that it enables comparison of targeting effectiveness across different interventions, contexts, and metrics.

The outreach ratio can be based on different measures. If, for example, the basis is poverty headcount, then the outreach ratio is the share of participants falling below the poverty line divided by the share falling below the poverty line in the entire population. If the

outreach ratio is above one, targeting is progressive, since the share of poor people participating is greater than in the population as a whole. If it is below one, targeting is regressive.

The key issue is to choose the basis of the outreach ratio. In doing that, we draw on the literature mentioned above, but we add a new type of outreach ratio inspired by the literature on poverty measurement, specifically Sen (1976) and Foster *et al.* (1984). In total, we will include outreach ratios based on four different poverty metrics, one of which we are the first to use in a targeting analysis.

A simple approach is to base the outreach ratio on levels of income or consumption as illustrated in the following equation:

$$OR_c = \frac{\frac{1}{N} \sum_{i=1}^n (y_i)}{\frac{1}{N_p} \sum_{i=1}^{n_p} (y_i)} \quad (1)$$

where OR_c is the outreach ratio based on consumption, N_p is the number of participants, N is the total number of observations, and y_i is the consumption of household i . Amin *et al.* (2003) do not construct the outreach ratio, but answer the same question by comparing incomes among participants and non-participants. In a situation where we want to reach the poor, the limitation of the outreach ratio based on income levels is that the rich contribute to the average to the same extent as the poor. An example of a hypothetical change in income illustrates the problem: A decrease in income among a rich participant, with everything else staying the same, would make targeting more progressive, even though no additional poor people are reached.

To overcome this issue, several studies construct the outreach ratio using poverty headcounts. The corresponding equation is:

$$OR_{ph} = \frac{\frac{1}{N_p} \sum_{i=1}^{n_p} (I_i(y_i < z))}{\frac{1}{N} \sum_{i=1}^n (I_i(y_i < z))} \quad (2)$$

where OR_{ph} is the outreach ratio based on poverty headcount, z is the poverty line and I_i is an indicator which is zero if household i 's income is above the poverty line, and one if it is below. Note that the metric for all participants is in the denominator, and not in the numerator as was the case in equation (1). This is to ensure that the ratio is above one when targeting is progressive. This exact ratio is used by Coady *et al.* (2004) and Handa *et al.* (2012). Mohammed *et al.* (1999) compare participation rates among the poor and non-poor, which provide similar insights into the outreach. The consumption levels of the non-poor households do not affect the outreach ratio based on poverty headcount. But there is another caveat, again illustrated by a hypothetical example: If we reduce the consumption level of a poor participant, the targeting metric remains the same, even though we now reach deeper than before.

This leads to a third metric, used by Park *et al.* (2002) in an analysis of Chinese counties, based on the so-called poverty gap measure, called the targeting income gap. The targeting income gap is the absolute distance from a county's average income to the poverty line summed over all mistargeted counties, i.e. counties that are in the program, but should not have been, and counties that are not in the program, but should have been. This measure is not subject to any of the critiques we discussed above: If a county with average income below the poverty line experiences a reduction in income and everything else stays the same, then the poverty gap measure increases. Applying this to individual level data, thus calculating the outreach ratio based on the poverty gap (OR_{pg}), is straight forward. Using the same notation as in equation (2), the equation is the following:

$$OR_{pg} = \frac{\frac{1}{N_p} \sum_{i=1}^{n_p} (I_i(y_i < z)(z - y_i))}{\frac{1}{N} \sum_{i=1}^n (I_i(y_i < z)(z - y_i))} \quad (3)$$

One way of comparing the outreach ratios based on the poverty headcount (equation 2) and the poverty gap (equation 3) is how they weight people below the poverty line. Poverty headcount assigns a weight of one to everyone below the poverty line and a weight of zero to households above the poverty line. The poverty gap uses the distance to the poverty line as weights. One critique of this weighting scheme is that it ignores the depth of poverty in the sense that an increase in income counts the same no matter how poor the household is, as long as it is under the poverty line (Foster *et al.* 1984). The hypothetical example illustrating this problem is as follows: A transfer from a poor participant to a richer participant, still under the poverty line, and where everything else stays the same, would leave the measure unchanged.

Using the squared poverty gap as a basis for the outreach ratio overcomes this critique. Park *et al.* (2002) suggest this but do not implement it. To the best of our knowledge, this is in fact the first use of the squared poverty gap as a basis for a targeting metric. Given the popularity of the measure in poverty analysis (Foster *et al.* 2010), this is peculiar. The outreach ratio becomes:

$$OR_{spg} = \frac{\frac{1}{N_p} \sum_{i=1}^{n_p} (I_i(y_i < z)(z - y_i)^2)}{\frac{1}{N} \sum_{i=1}^n (I_i(y_i < z)(z - y_i)^2)} \quad (4)$$

Again, the notation follows equation (2). Since the different outreach ratios give different weights to the poor, we implement all four in our analysis of targeting effectiveness.

In assessing the metrics, we have implicitly judged them using critiques often raised in the literature on poverty measurement. As is commonly done in this literature, we can now sum this up as three

principles that a targeting metric must meet, assuming that the program targets the poor:

- The threshold principle: A change in the income of a person above the poverty line should affect the targeting metric less than a change in the income of a person below the poverty line.
- The poverty principle: A reduction in income among participants below the poverty line must make the targeting metric more progressive.
- The distribution principle: A transfer of income from a participant below the poverty line to any participant who is richer must make the targeting metric more progressive.

The four different outreach ratios discussed above are particularly appropriate for assessing targeting based on a continuous variable, such as total consumption. But we also compute outreach based on two simpler consumption metrics, specifically meals per day and the length of the hungry period. To use the outreach ratio requires defining a poverty line and measuring the distance to this poverty line for each household. We define poverty lines in the section on empirical strategy. Finally, we also include variables on education and health to analyze multidimensional poverty. For these we calculate only outreach ratios based on levels similar to OR_c in equation (1).

After computing the outreach ratios, we present some self-reported reasons for why some people do not join. The survey leaves us with very little data on this issue, but we will nevertheless provide some descriptive results since this issue is important for the usefulness of the results.

Table 1. Overview of the literature on targeting used in the text.

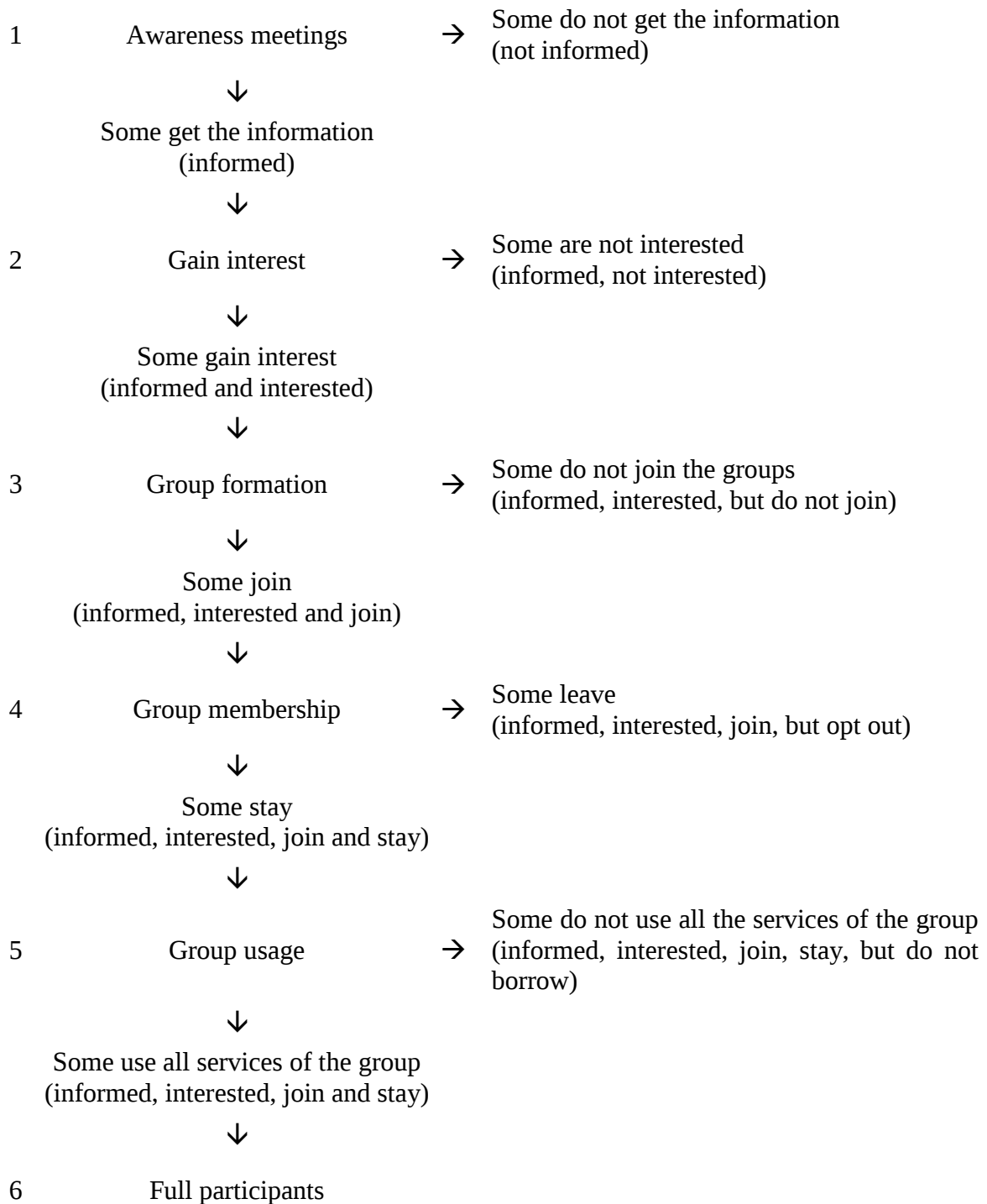
	Country	Interventions analyzed	Method	Dimension	Findings	Endogeneity issues
Microfinance						
Amin <i>et al.</i> (2003)	Bangladesh	Grameen Bank, ASA, BRAC	Compares consumption and poverty levels for participants and non-participants.	Poverty levels	Participants are poorer than non-participants.	No. Surveys 1992, use membership 1995.
Mohammed <i>et al.</i> (1999)	Bangladesh	BRAC	Compares participation rates across three wealth groups	Participation rates and poverty levels	The poor participate more than the non-poor	Yes. Survey from 1994 only.
Navajas <i>et al.</i> (2000)	Bolivia	Several microfinance institutions	Analyze poverty level compared to national averages	Poverty levels	Participants are just below the poverty line, not poorer.	No
Other						
Coady <i>et al.</i> (2004)	Multiple countries	111 targeted anti-poverty programs	Divides the share of transfers to a segment of a population with the segment's share of the total population	Transfer of resources	The median intervention is progressive, but a quarter is regressive.	Not certain.
Ravallion (2009)	China	Di Bao transfer program	Compare different targeting metrics' ability to predict differences in poverty reduction due to transfers	Participation rates	Typical targeting metrics do not predict poverty reduction impact	Yes. Assumes pre-program income to be current income minus transfer.
Handa <i>et al.</i> (2012)	Malawi, Kenya,	Social transfer programs	Divides the proportion of participants in the lowest quintile with the proportion of the population in the lowest quintile (20%).	Participation rates	The poor participate more	Not relevant. Compare participants to national averages only, not non-participants in the same area
Park <i>et al.</i> (2002)	China	Various interventions	Unit of analysis is the county and their eligibility for transfers. Measure is the proportion of mis-targeted counties and a measure of the extent of mistargeting	Participation rates	High degree of mistargeting: 22% of counties	No
Houssou and Zeller (2011)	Malawi	Agricultural input subsidies in Malawi	Analyze the performance of alternative targeting metrics	Participation rates	An indicator-based metric is superior.	Not relevant. Compares different metrics only

A Leaking Pipeline

As is clear from the previous sections, many studies have looked at who is reached by microfinance, including the poverty status of these. There have been several examples where microfinance has failed to reach the poor. A natural question is: Why are the poor not included? What mechanisms lead to non-participation of the poorest in microfinance and what can be done to prevent this from happening? To investigate this, we borrow the metaphor of a leaking pipeline, which has been used in the literature on gender disparities in education and academia (Barinaga 1992, White 2004). The leaking pipeline illustrates the fact that women exit the academic career path on several steps along the way to professorship. This type of sequential approach has been used to a lesser extent in studying social programs in developed countries (Heckman and Smith 2004) and only recently in developing countries as well (Coady *et al.* 2013).

We identify five steps where households can exit the pipeline of participation (see table 2 for an overview). First, the poorest might not receive the information about the upcoming awareness meetings and hence do not know about the groups being initiated. Information about the project is likely to spread via informal channels in addition to the actual meetings, but it is possible that the poorest are excluded from these information networks as well. Second, of all those who get the information, only some turn out to be interested. Some people might not need the services of the groups, or they might already at this stage think that they will be unable to find a group which will accept them. Third, the group formation in itself might leave some out, even if they are interested, as it happens voluntarily—villagers are asked to form groups with other people from the village whom they trust. It could be that assortative matching takes place such that riskier borrowers join groups together, following the traditional theory of information asymmetry (Ghatak and Guinnane 1999). If poverty status is correlated with risk aversion, this might affect the outreach, since only a limited number of groups are formed in each village. Alternatively, poorer households could be seen as less advantageous for a group and therefore will have difficulties finding one.

Table 2. The pipeline



Once the groups have been formed, some attrition will eventually happen, which is the fourth step. It might be the case that the poorest do not find the groups useful or that they are pressurized into leaving

the group by other members. Finally, as a fifth step, even if the poorest stay in the groups, it is possible that they do not use all the services of the group. Possibly, the poorest do not have enough non-credit resources to make use of loans, or their income is too volatile to risk an involuntary default making them focus on smoothing income instead (Morduch 1995).

Following these stages where exiting is possible, we arrive at six mutually exclusive groups: One for each stage of exiting and a sixth group containing full participants. We compare the households that exit with those who stay in the pipeline with respect to poverty levels using the metrics proposed in the previous sections.

The Data

The data were collected in one sub-district of Karonga in northern Malawi during two six-week periods in August 2009 and August 2011 as part of the randomized controlled trial documented in the first paper of this thesis. A total of 3,700 households with 20,800 people live in the villages, which cover an area of approximately 400 square kilometers.

The total sample consists of 890 households from twenty-three villages. The entire survey covered forty-six villages, but VSLAs were only established in a randomly selected half of the villages so the current analysis only uses data from the villages in which the VSLA groups were established.

Interviewed households were sampled randomly from household lists provided by local authorities. Stratified sampling was done using two criteria: Village and initial household interest in participating in the VSLAs. In total, data were collected from forty-six villages with interested and non-interested households in each, i.e. a total of ninety-two strata, but since we use only data from half of the villages, we have 46 strata altogether. A higher propensity of sampling was chosen for households in smaller villages relative to larger villages as well as for households who initially expressed interest relative to households not showing interest. Village stratification was performed because of the village-level randomization in the impact assessment.

Oversampling of interested households had the aim of oversampling final participants in treatment villages and potential participants in control villages, also for the purpose of assessing impact. Results reported below are weighted according to the inverse probability of sampling following standard practice in survey research and as recommended by several authors, for example Deaton (1997) and Solon *et al.* (2013).

The randomization of villages into treatment and control groups was done within seven blocks to increase baseline balance. The variables used in the analysis below are from the baseline survey in 2009 except for the variable indicating participation in a VSLA, which was determined in the follow-up survey in 2011. The attrition rate from 2009 to 2011 was less than three percent, which is low compared to similar surveys (Glewwe and Jacoby 2000). In total, forty-eight percent of households participated in VSLAs in 2011.

The questions allowing us to later identify the pipeline were not asked to all respondents but only to a random half using a questionnaire which was longer than the standard questionnaire. Below, this questionnaire is referred to as the long questionnaire, as opposed to the standard questionnaire, which contained only a subset of the questions in the long questionnaire. The purpose of the long questionnaire was to gather information particularly on future participants, and thus households who indicated interest in participating were oversampled to a larger extent than in the short questionnaires. This leads to a different weighting scheme when we analyze the long questionnaire data only. On the other hand, the standard questionnaire increased the power of detecting an effect for the impact assessment in terms of participants and non-participants alike. For these reasons our sample size decreases when we discuss the pipeline compared to simply looking at whether the poor households participate in the VSLA groups.

We use four consumption measures in what follows: Total consumption calculated using recall questions regarding the consumption of seventeen food items in the past week, total consumption predicted using USAID's Poverty Assessment Tool for

Malawi (PAT), meals per day, and finally the length of the hungry period measured as the number of months where the household consumed less than three meals a day.

For the first measure, seventeen food items were identified from the Malawi Second Integrated Household Survey from 2004/5 (IHS2). The seventeen items are the ones consumed the most in rural Karonga, the district of the survey, and made up eighty-nine percent of total food consumption and fifty-five percent of total consumption. The total consumption figure below is the total value of these seventeen items divided by fifty-five percent. The method used in summarizing the 17 items, in particular the calculation of the prices, is explained in detail in Appendix A in the first paper of this thesis. This appendix also describes some of the limitations of this measure, for example that we were forced to calculate our own conversions for each combination of items and units, e.g. a tin of maize, since the conversions provided in the IHS2 were unreliable. This can lead to measurement error. A common finding is that poorer households spend a larger share of their income on food compared to richer households, and thus one concern when using this method is whether it overestimates total consumption for the poor while underestimating it for the rich. In our case, a regression of the share of consumption spent on food on total consumption using the 2004 integrated household survey data from the area, results in a negative, but insignificant estimate ($t=-1.24$).

The second consumption measure is USAID's Poverty Assessment Tool (PAT), which is 20 questions selected on the basis of their ability to predict total consumption in data from the Malawi Second Integrated Household Survey (IRIS Center 2012). We included these questions in our survey and use them and the parameters provided by USAID PAT to predict total consumption for each household. All USD figures are in 2005 dollars using the exchange rate of 91 MKW/USD, adjusted for inflation and using the poverty-adjusted purchase power parity exchange rate described in Deaton and Dupriez (2011).

Apart from the two consumption measures described above, we include two simpler consumption measures: “Meals per day” and “length of the hungry period”. “Meals per day” is simply the number of meals the household consumed the day before the survey, and the hungry period is the number of months within the last year where the household consumed less than three meals per day using recall. This occurs most often in the period just before the green harvest in April. However, the length of the hunger period differs greatly among households. While these measures are cruder than the two consumption measures mentioned earlier, they are easier to measure and thus may contain less measurement error.

Finally, we include four indicators of education and health, which are components commonly included in multidimensional poverty (Alkire and Foster 2011). For education, we look at years of schooling for the household head as well as the share of children aged sixteen to twenty-five who are in school. We choose this age group since this particular area of Malawi is known for a general high level of primary education, and we therefore do not expect much variation for other ages. For health, we use a subjective health measure indicating whether each individual’s health is very good, good, average, bad, or very bad. We include both the household average and an indicator for households having one or more members in “bad” or “very bad” health. raising livestock, or fishing.

Table 3 shows summary statistics. The top rows of the table show measures of consumption and food security. Fifty-one percent of the population lives below the 1.25 USD poverty line and one third eat less than three meals per day in August, which is just after the harvest that occurs between May and July. The average hunger period is four months. The middle rows display household characteristics. An average household had almost six members and 17% were headed by women. The last part of the table shows the livelihood patterns with 80% being involved in farming, twenty-one percent in fishing, and sixty-two percent in some income generating activity other than agriculture, raising livestock, or fishing.

Table 3. Descriptive statistics

	Mean	SD
Consumption (USD/capita/day, log)	0.20	0.48
Household living below the 1.25 USD poverty line	0.51	0.44
Meals per day	2.61	0.49
Household consumed less than three meals yesterday	0.36	0.42
Hungry period in months	4.24	3.58
At least one hungry month during the last year	0.74	0.38
Number of household members at time of interview	5.71	2.05
Household head is a women	0.17	0.33
Years of education of household head	7.06	2.82
Household health score (1=good)	1.62	0.50
Anyone in the household has bad health	0.11	0.27
Share of children age 16-25 currently in school	0.10	0.17
Household member of VSLA group in 2011	0.44	0.43
Household does petty trade or small business	0.55	0.43
Household does fishing	0.21	0.36
Household does any farming (subsistence, cash crop, or livestock)	0.80	0.35
Any income-generating activities (excluding agriculture and livestock)	0.63	0.42
Agriculture is the most important income source	0.54	0.43
Land ownership in acres	2.48	1.80
Number of rooms	2.73	1.07
Number of observations	874	

Note: The number of observations is lower than in the main analysis due to missing observations on the age of household head. All statistics are on the same sample and computed using sampling weights.

Empirical Strategy

As mentioned in the section on targeting measures, our primary measure of targeting effectiveness is the outreach ratio which compares the poverty status of participants – irrespective of whether they use the loan feature or not – to the poverty status of the population as a whole. In this way, it compares the actual targeting with neutral targeting, i.e. the situation where households participate irrespective of their poverty status.

In effect, the outreach ratios are all ratios of averages taken over four corresponding household metrics in the participant group and

among all households, respectively. The four household level poverty metrics are:

Consumption (used in OR_c):	$pm_i^c = y_i$
Poverty headcount (used in OR_{ph}):	$pm_i^{ph} = I_i(y_i < z)$
Poverty gap (used in OR_{pg}):	$pm_i^{pg} = I_i(y_i < z)(z - y_i)$
Squared poverty gap (used in OR_{spg}):	$pm_i^{spg} = I_i(y_i < z)(z - y_i)^2$

where, as above, y_i is consumption, I_i is an indicator which equals one if the household is poor and zero otherwise, and z is the poverty line. For the measures based on 2005 USD we follow Chen and Ravallion (2010) in choosing 1.25 USD as the poverty line. For the simple measures there are levels which can be considered a natural choice in identifying the poor. Regarding meals per day we classify households consuming two meals or less as poor. For hungry months, we label households as poor if they indicate one or more hungry months. The distance to the poverty line is then simply the number of hungry months.

When it comes to the multi-dimensional poverty measures, there is no agreement on how to sum up across the different measures. One attempt was made by Alkire and Foster (2011) and used in the Human Development Report 2010 (UNDP 2010), but this approach was later criticized particularly because the authors add up different measures with arbitrary weights (Ravallion 2011). Because of this, we do not sum up the multidimensional measures but simply display them one by one.

As already discussed, the issue is whether or not the outreach ratio is different from one. An outreach ratio smaller than one corresponds to regressive targeting, whereas a ratio larger than one is progressive targeting. To assess if the ratio is indeed statistically significantly different from unity, we compare the average of the participants with the average of non-participants by fitting the following very simple regression model:

$$pm_i^k = \alpha + \beta D_i + \varepsilon_i \quad (5)$$

where pm_i^k is the individual poverty metric z as listed above for household i , D_i is a dummy which is equal to one when anyone from the household participates in a VSLA in 2011, and ε_i is the error term. We also estimate the total mean, $\mu = \frac{1}{N} \sum_{i=1}^N (w_i * pm_i^k)$ where w_i is the sampling weight of household i . The average level for participants is $\alpha + \beta$. We assess whether α/μ is different from one by testing $H_0: \beta=0$, which is effectively a weighted t-test for equality of means between the two groups. We always use the inverse probability of sampling as weights as explained in the first paper of this thesis.

When investigating the pipeline, we use a simple version of the approach used by Heckman and Smith (2004), who analyze participation in the job training program JTPA as a pipeline. As mentioned above, we analyze leakages in five sections of the pipeline, which can be formulated as conditional probabilities: (1) the probability of gaining awareness of the project, (2) the probability of interest given awareness, (3) the probability of joining a group given interest and awareness, (4) the probability of staying in a group, i.e. not dropping out, given joining a group, awareness and interest and (5) the probability of utilizing the full range of services provided by the group, i.e. both loans and savings given (1) to (4). More formally, we have:

(6)

$$P_1 = \Pr(\text{aw}=1)$$

$$P_2 = \Pr(\text{int}=1 \mid \text{aw}=1)$$

$$P_3 = \Pr(\text{join}=1 \mid \text{int}=1, \text{aw}=1)$$

$$P_4 = \Pr(\text{stay}=1 \mid \text{join}=1, \text{int}=1, \text{aw}=1)$$

$$P_5 = \Pr(\text{use}=1 \mid \text{stay}=1, \text{join}=1, \text{int}=1, \text{aw}=1)$$

To compute these probabilities we use data from several sources. Steps one and two draw on the baseline survey. Steps three, four, and

five use data from the follow-up survey two years later. Apart from estimating the actual probabilities, we also estimate average poverty metrics within each of the five nested groups. We test if there is a difference between the group that stays in the pipeline and the one that exits. We do not calculate the outreach ratio since we do not compare the groups to the entire population. For example, to test for difference in poverty among those who join and those who do not join a group is defined as the following conditional expectation:

$$E(y_i | \text{join}=1, \text{int}=1, \text{aw}=1) - E(y_i | \text{join}=0, \text{int}=1, \text{aw}=1) \quad (7)$$

We test the H_0 that this expectation is equal to zero using the same regression model as (5) above, where D_i in this case is a dummy for joining a group, and the only households included are the ones who were aware and interested. A natural challenge in doing this is that the sample size, and thus the power, decrease throughout the pipeline. The chance of not rejecting H_0 , even when it is in fact false, increases.

Results

There are clear signs of regressive targeting across almost all the different poverty metrics and the four different outreach ratios described in our analysis above as displayed in table 4. The only notable exception is when the outreach ratio is based on directly measured consumption calculated from 17 food items, in which case we find that the outreach ratio is not different from one.

When it comes to the outreach ratio based on the USAID PAT, it is lower than one, and the difference is significant at a ten percent level for ratios based on levels, poverty gaps, and squared poverty gaps. Using the PAT consumption level results in a ratio of 0.8, since log of consumption is 0.16 among participants and 0.20 in the whole population. This difference of 0.04 corresponds to a 4% difference in consumption levels or approximately 5 cents, using the formula

provided in Kennedy (1981).⁴⁴ The outreach ratios based on poverty gap and squared poverty gap are 0.85 and 0.90, respectively, indicating a 10-15% difference between participants and the whole population in these poverty metrics. Since results are significant only at a ten percent level, we interpret this merely as indications of differences.

The results on the simpler measures are stronger, however. Within these the outreach ratios are consistently below one, meaning that targeting is regressive, and participants are systematically better off compared to the population as a whole. For the length of the hungry month and meals consumed yesterday results are significant at 1% or 5% levels, except for the poverty headcount with regard to the hungry period. This is one limitation of setting the “poverty line” at zero, since it leads to almost all households being poor, and there is thus little difference in the poverty levels of participants and non-participants. However, the outreach ratios that measure the distance to the created poverty line overcome this problem and find significant differences.

Interestingly, the outreach ratio worsens as we apply greater weights to the poorest. The outreach ratio based on the average number of meals per day is 0.97. Based on the share of households eating less than three meals it is 0.86. Looking at the poverty headcount, poverty gap and squared poverty gap it is reduced to 0.86, 0.79, and 0.68, respectively, all significant at a 1% level. An illustration of this is provided in figure 2, which shows that almost none of the households consuming one meal per day are members. The hungry period follows the same pattern, where participants face a period which is half a month shorter than the one the overall population experiences. Figure 3 displays the entire distribution and shows a clear difference among the poorest, for example that twice as many non-members as members have a constant hungry period, i.e. consume less than three daily meals in all twelve of the last twelve

⁴⁴ The formula is $\Delta y = e^{\left(\hat{\beta} \frac{\text{var}(\hat{\beta})}{2}\right)}$.

months. Among the households experiencing a relatively short hunger period – between one and four months – the members are over represented. This also explains why the outreach ratio based on the poverty headcount is not different from one: Since the poverty line is set at one hungry month, households with short and long hungry periods all count the same.

Regarding our measures of multidimensional poverty, we cannot reject neutral targeting with respect to health, but we find regressive targeting when it comes to children's education levels, where both indicators are statistically significant, albeit only at a ten percent level.

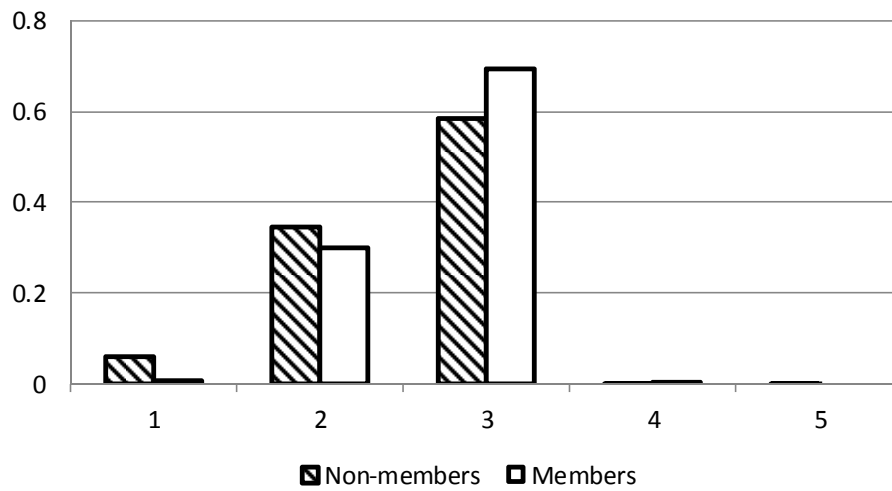
In sum, these results indicate that not even the most pro-poor microfinance can reach the poorest. As a final note, though, we compare the targeting effectiveness of VSLA to other providers of microfinance. Several established microfinance institutions operate in the area, for example FINCA, Pride Africa, Malawi Rural Finance, and Opportunity International, and 0.5% of the households in our sample have an account in these institutions. If we extend the pool to friends and relatives, which was the most common source of finance before the intervention, the percentage with loans is five percent. We interpret these results in the conclusion.

Table 4. Main results

Basis for calculating the outreach ratio	Outreach ratio	Averages	
		Participants	All households
Consumption poverty: Directly measured			
Consumption (USD/capita/day, log)	1.10	0.18	0.20
Poverty headcount	1.03	0.51	0.50
Poverty gap	1.02	0.17	0.17
Squared poverty gap	0.98	0.08	0.08
Consumption poverty: Indirectly measured			
Consumption using PAT (USD/capita/day, log)	0.80	0.20	0.16
Poverty headcount using PAT	0.94	0.52	0.55
Poverty gap using PAT	0.90*	0.15	0.16
Squared poverty gap using PAT	0.84*	0.05	0.06
Consumption poverty: Simple measures			
Meals per day	0.97**	2.69	2.61
Poverty headcount (less than three meals per day)	0.86**	0.31	0.35
Poverty gap (meals)	0.79***	0.10	0.13
Squared poverty gap (meals)	0.70***	0.04	0.05
Hungry period in months	0.89**	3.76	4.20
Poverty headcount (at least one hungry month last year)	0.99	0.72	0.73
Poverty gap (hungry period)	0.89**	3.76	4.20
Squared poverty gap (hungry period)	0.82***	28.44	34.69
Multidimensional poverty			
Years of education of household head	0.96	7.34	7.07
Share of children age 16-25 currently in school	0.81*	0.12	0.09
Anyone in the household has bad health	0.95	0.11	0.11
Household health score (1=good)	0.97	1.58	1.62
Observations	883		

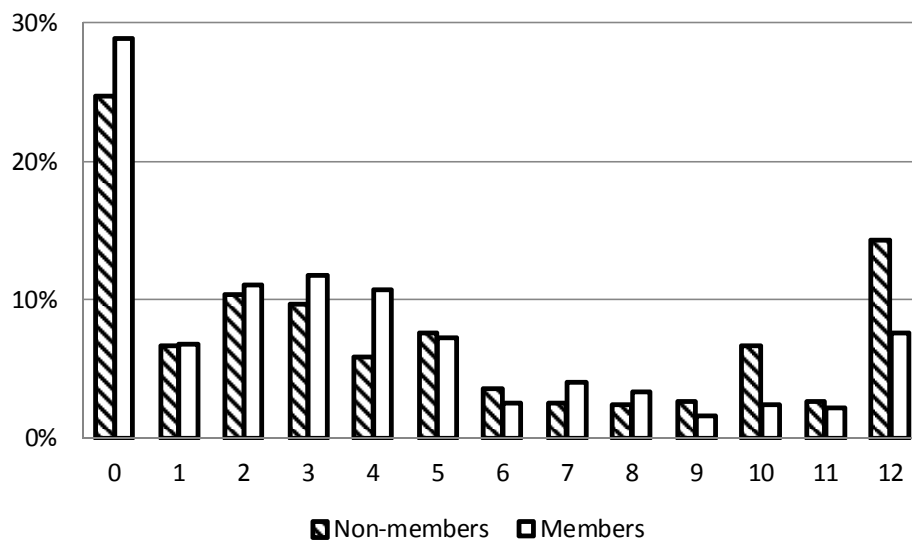
Note: This table shows that there is regressive targeting in VSLA. Stars on the outreach ratio are a result of a t-test of whether there is a difference among participants and non-participants, and thus whether or not the outreach ratio is different from one. The test is performed as a regression of the poverty parameter on a dummy for non-participation where $H_0: \beta=0$. * Significant at 10%, ** Significant at 5%, *** Significant at 1%. All estimations use sampling weights and robust standard errors. Clustering standard errors at the village level yield similar results.

Figure 2. Food consumption (meals/day)



Note: Proportions are estimated using sampling weights and sum to 100% across non-members and members.

Figure 3. Hungry period in months



Note: Proportions are estimated using sampling weights and sum to 100% across non-members and members.

What Prevents the Poor From Joining?

Our results from the previous section show that regressive targeting takes place in the intervention. But what are the reasons for this? Since there is self-selection into the groups, the issue is whether the poor are left out because the other members do not want them, or because they themselves do not want to join a group. In the latter case, why do they not want to join? Due to the importance of the issue, we give some preliminary answers to these questions below before investigating the participation pipeline described earlier. However, the following section draws on data from only seventy-three households.

The data we use come from an additional survey administered halfway between the baseline and the follow-up, i.e. after one year, which is the only time we asked questions regarding this issue. Two selection criteria make the sample size small: We ask only respondents who are not already members of a VSLA (564/801), and we limit the questions to the non-members who, at the time before start-up, actually had knowledge that groups were starting up (73/564). Eight of these seventy-three initially joined groups but dropped out, and further thirteen were not interested in joining. The remaining fifty-two were interested, but did not join. The reasons stated by all seventy-three are summarized in table 5.

The primary reason is lack of demand with 34.7% answering “Didn’t have money to save”, followed by lack of supply: “Not enough groups formed” (15.1%). The latter could both reflect a lack of supply from the NGO or the inability of the household to find a group willing to let the household join. This indicates that the intervention design itself might prevent some people from participating. Unfortunately, the sample size is too small to investigate whether the poorest fall in either of these categories.

Practitioners and academics have repeatedly refuted the claim that someone can be too poor to save (Bouman 1984, Armendariz de Aghion and Morduch 2005, Rutherford 2009). Given the responses above, this is not as straightforward as these authors claim. If by savings we understand cash savings, then some people might actually have too little cash to save, and VSLA is not suitable for them.

Another explanation could be that the dynamics of group formation leads to groups with too high minimum savings requirements at least in some villages. Following this strain of thought, VSLA in and of itself might be suitable even for the poorest, but the way the groups are formed does not ensure that groups can accommodate small savings.

Table 5. Reasons for not joining a VSLA group

No need for the services in the group	6.6%
Other members in the group did not want me to join	5.0%
I didn't know time/location of meetings	3.5%
Not enough groups formed	15.1%
My husband did not approve	9.3%
Didn't have money to save	34.7%
Temporarily away at the time	13.1%
Other	12.6%
Total	100.0%
N	73

Note: The table show that lack of money to save is an important constraint preventing people from joining VSLAs. Respondents are people who knew VSLA was starting before it started and therefore 73. Control village with VSLAs are included. Proportions estimated using sampling weights.

Investigating the Pipeline

We now turn to the pipeline of participation. As mentioned in the section on data, the analysis includes only half of the respondents as opposed to the main analysis, and as we move down the pipeline the number decreases.

Table 6 shows how membership develops through the pipeline. The first column shows the conditional probabilities mentioned in equations (6)a to (6)e.

Table 6. Pipeline results

Section of the pipeline	Conditional probability		Share of total in the pipeline	Share of total leaving the pipeline
	$P_x = \Pr(j=1 \mid \text{previous steps})$			
1) Gain awareness	60%		60%	40%
2) Express interest	88%		53%	7%
3) Join group	59%		31%	22%
4) Stay in group	95%		30%	2%
5) Use both savings and loans	64%		19%	11%

Note: The table shows that only 60% of households got information about the intervention prior to start-up and no more than 59% of those who got information and were interested, actually joined. All percentages are estimated using sampling weights.

Two interesting facts emerge from the table: The awareness campaign reached sixty percent of the population. If the ambition is to reach everyone, and considering that the area is fairly small, then this is low. Moreover, even though almost all of those who hear about the project are interested, only fifty-nine percent of these choose to join a group. For some reason, forty-one percent of those who hear about the project *and* express interest end up not joining two years down the line. In order to understand how participation happens, this is important. The flip side of this is that a large share of those who initially said they were not interested ended up joining. Awareness campaigns are needed, but in their current design they are not necessarily good at communicating costs and benefits of participating in ways that match the intervention.

After looking at the general pipeline, we now analyze the poverty profile of the pipeline. We start out by looking at actual participation on this smaller sample available for the analysis of the pipeline. Since the questions regarding the pipeline were administered to only half of the households, there may be differences between this randomly chosen half of the sample and the full sample. To investigate this, we perform the general participation analysis on this subset of households (the first column in table 7). The results from the general analysis disappear, and in one case—that of directly measured consumption—

we find a significant result where the main analysis found no significant difference. Apparently, although the sample was randomly chosen, there are differences between the whole sample and the subset of the sample that was given the long questionnaires, although there are no cases where the results have opposite signs and are statistically significant at the same time.

Going into details in the different elements of the pipeline we estimate the regression model in equation (5) for each of the steps in the pipeline. This tells us if those in the pipeline differ from those exiting the pipeline. The results are shown in table 7, columns 2 to 6. In most of the steps there are no systematic differences between those who exit and those who stay in the pipeline. There is, however, one important difference: The people who join a group, conditional on getting information about the project and indicating awareness, are significantly poorer than those who do not join. The households who join have lower average consumption, higher poverty headcount, and higher squared poverty gap, indicating that even when looking at the very poorest and taking the distribution of poverty into account, there is a difference.

Since we find no difference on the awareness campaign, we conclude that information about the project reaches a wide, and poverty-wise representative, selection of households, but that the actual activities initially attract the less well-off and appeal less to the richer. Also, the richer join later since we find the opposite difference in the analysis of final participation above—a result of what has been termed “injections” into the pipeline (Soe and Elaine 2008). One reason for this, which is sometimes voiced by practitioners, is that in areas with aid dependency, the well-connected are attracted by the expectation that they can receive transfers of money or goods. Since VSLA does not involve transfer of resources to households, the richer households, who first expressed interest, might be discouraged from spending the time group membership takes.

For the same reason, these results do not support a common assertion by practitioners, i.e. that the poorest are at first reluctant and then join later, possibly because they are risk averse and afraid to

place the little money they have outside of the home. If anything, the results point in the opposite direction: The poorest join first.

Table 7. Poverty profile of the pipeline

	(1) Membership (long q only)			(2) Gain awareness			(3) Express interest			(4) Join group			(5) Stay in group			(6) Use both savings and loans		
	Pipel	Diff	t	Pipel	Diff	t	Pipel	Diff	t	Pipel	Diff	t	Pipel	Diff	t	Pipel	Diff	t
Consumption poverty: Directly measured																		
Consumption (USD/capita/day, log)	0.09	0.20	2.15**	0.20	0.01	0.17	0.21	-0.07	-0.50	0.10	0.25	2.51**	0.10	0.11	0.72	0.13	-0.09	-0.68
Poverty headcount	0.58	-0.14	-1.31	0.47	0.06	0.79	0.46	0.15	1.24	0.54	-0.20	-2.18**	0.54	-0.06	-0.34	0.48	0.17	1.15
Poverty gap	0.20	-0.06	-1.48	0.17	-0.01	-0.34	0.16	0.06	1.01	0.19	-0.07	-1.48	0.19	-0.03	-0.35	0.17	0.06	0.96
Squared poverty gap	0.09	-0.03	-1.41	0.08	-0.01	-0.73	0.08	0.01	0.40	0.09	-0.03	-0.95	0.09	-0.03	-0.83	0.09	0.01	0.14
Consumption poverty: Indirectly measured																		
Consumption using PAT (USD/capita/day, log)	0.10	0.06	0.76	0.12	0.04	0.64	0.12	0.00	0.01	0.04	0.20	2.38**	0.05	-0.05	-0.45	0.07	-0.07	-0.58
Poverty headcount using PAT	0.59	0.07	0.75	0.60	0.00	0.00	0.59	0.07	0.51	0.63	-0.11	-1.16	0.63	0.09	0.54	0.58	0.14	0.98
Poverty gap using PAT	0.19	-0.04	-1.06	0.17	-0.01	-0.35	0.17	-0.04	-1.05	0.21	-0.09	-2.27**	0.21	-0.01	-0.13	0.20	0.03	0.45
Squared poverty gap using PAT	0.07	-0.02	-1.03	0.06	0.00	-0.25	0.07	-0.03	-1.95*	0.09	-0.05	-2.23**	0.09	-0.02	-0.49	0.09	0.00	-0.02
Consumption poverty: Simple measures																		
Meals per day	2.72	0.04	0.39	2.77	-0.03	-0.45	2.77	-0.01	-0.04	2.71	0.15	1.77*	2.71	-0.07	-0.38	2.72	-0.03	-0.16
Poverty headcount (less than three meals per day)	0.28	-0.03	-0.35	0.24	0.02	0.25	0.24	-0.01	-0.07	0.30	-0.14	-1.59	0.29	0.07	0.36	0.28	0.03	0.14
Poverty gap (meals)	0.09	-0.01	-0.33	0.08	0.01	0.39	0.08	0.00	0.06	0.10	-0.05	-1.59	0.10	0.02	0.36	0.10	0.01	0.14
Squared poverty gap (meals)	0.03	0.00	-0.29	0.03	0.01	0.63	0.03	0.00	0.30	0.03	-0.02	-1.59	0.03	0.01	0.36	0.03	0.00	0.14
Hungry period in months	3.54	0.08	0.09	3.52	-0.04	-0.06	3.43	0.76	0.88	3.57	-0.33	-0.42	3.54	0.42	0.31	3.20	0.96	0.53
Poverty headcount (at least one hungry month last year)	0.77	-0.03	-0.3	0.82	-0.11	-1.47	0.82	0.03	0.38	0.81	0.01	0.13	0.81	0.04	0.32	0.84	-0.08	-0.65
Poverty gap (hungry period)	3.54	0.08	0.09	3.52	-0.04	-0.06	3.43	0.76	0.88	3.57	-0.33	-0.42	3.54	0.42	0.31	3.20	0.96	0.53
Squared poverty gap (hungry period)	26.45	-0.31	-0.03	23.84	3.40	0.49	23.09	6.20	0.59	25.72	-6.39	-0.63	25.50	4.28	0.24	19.22	17.55	0.72
Multidimensional poverty measures																		
Years of education of household head	7.17	-1.02	-1.36	7.09	-0.55	-0.99	7.09	0.03	0.04	7.13	-0.09	-0.14	7.25	-2.30	-3.08***	7.14	0.30	0.35
Share of children age 16-25 currently in school	0.13	-0.07	-1.52	0.10	0.01	0.18	0.11	-0.06	-1.93*	0.12	-0.04	-0.87	0.13	-0.05	-0.83	0.09	0.11	1.39
Anyone in the household has bad health	0.06	0.04	0.9	0.05	0.06	1.55	0.06	-0.03	-0.76	0.04	0.05	1.68*	0.03	0.08	1.02	0.05	-0.05	-2.32**
Household health score (1=good)	1.61	0.07	0.59	1.55	0.26	2.69***	1.54	0.08	0.62	1.58	-0.11	-1.06	1.58	0.06	0.34	1.61	-0.08	-0.50
Observations	414			414			251			221			128			117		

Note: This table demonstrates that there is difference in poverty levels among those who join VSLA and those who do not, when the analysis is made on the group who got information and expressed interest (column 4). Compared to the main analysis, only half of the observations are included since the information used was in half of the questionnaires, which were longer than the remaining. Column (1) shows that overall results on this part of the sample diverge from the main analysis with respect to directly measured consumption. "Pipel" is the value for the households still in the pipeline. "Diff" is the value for the households not in the pipeline minus the value for the households still in the pipeline. "t" is the t-value of the difference. * denotes statistical significance at the 10% level, ** denotes statistical significance at the 5% level, and *** denotes statistical significance at the 1% level. All estimates are calculated using sample weights.

Conclusion

Developing country governments, donors, and NGOs all want to reduce poverty through interventions that reach the poorest. Within microfinance, one way of achieving this goal has been to develop and implement community-managed methods particularly suited for people living on less than the ubiquitous “dollar a day.” Out of the 207m clients in microfinance worldwide, at least two million are members of some 87,000 savings groups similar to the ones we analyze in this paper (Maes and Reed 2012 and savingsgroups.com).

Our goal in this paper is to assess whether these methods are indeed successful at reaching the poorest. The answer is provided by analyzing panel data from one typical community-managed intervention in northern Malawi. Overall, we find regressive targeting: The participants are not as poor as the average population in the area.

Before looking at the data, we review the literature and identify the outreach ratio as a common and useful way of analyzing targeting. The outreach ratio is simply the share of poor people reached divided by the share of poor people in the population. As an extension of this, we develop our own outreach ratio based not on poverty headcount but on the squared poverty gap—a common metric in poverty measurement, which takes the extent and severity of poverty as well as the income distribution among the poor into account. No previous targeting metric has to our knowledge taken all these three issues into account. Our suggested metric does. Finally, we build a framework that allows us to investigate the participation decision sequentially—an approach we label the participation pipeline. The analysis thus serves not only as an analysis of targeting effectiveness in a particular intervention, but also as an application of these methodological developments.

Turning to the data, we find that fifty-five percent of the general population fall below the 1.25 USD PPP poverty line using USAID’s Poverty Assessment Tool, whereas fifty-two percent of the participants are poor using this definition. The outreach ratio is below one in almost all cases, which means that targeting is regressive. Using our newly developed metric, which gives higher weights to the

poorest, we find the same results: participants are not as poor as the average household in the area. The only exception is consumption measured directly using recall questions on seventeen food items, where we find no significant differences, and we provide some suggestions as to why that might be the case. Importantly, the results show the usefulness of our newly developed metric, since regressive targeting is even more widespread here compared to conventional metrics.

To study why the poor end up participating less than the non-poor, we look at reasons for not joining groups. Asking people why they do not join groups reveals that the savings requirements might be one reason. This is in opposition to the often stated claim that nobody is too poor to save. When investigating the participation pipeline, we find that the awareness campaign reaches less than two-thirds of the population in the area and that only about forty percent of those initially indicating interest end up joining. Somehow the awareness campaign is out of touch with the intervention.

As for the poverty profile through the pipeline, the awareness campaign reaches poor and non-poor alike, but out of those who get the information, the poorest join first. Only later do the non-poor join. For practitioners the last fact is likely to be new. A common assertion is that the poorest join later, whereas this points to the opposite.

What are the practical consequences of these findings? On the one hand, the results suggest that microfinance cannot, in fact, reach the poorest. The glass is half empty. On the other hand, however, a large fraction of the poorest does participate, and VSLA certainly reaches a much poorer group than conventional microfinance, which serves 0.5% of the population in the area. In other words: The glass is half full.

Whether or not we should adjust the VSLA model to reach poorer groups depends on local trickle down, i.e. whether or not non-participants benefit from VSLAs in their village even though they do not participate. If non-participants can benefit, then the current model could be sufficient. There is very limited knowledge on this, so it is a topic for future research. Implementation manuals could also be

focused on creating benefits in the local community beyond participants, for example through community engagement by groups. This could, however, affect the performance of the groups.

If there is no or little scope for local trickle down, then real targeting methods are needed in order for VSLAs to benefit the poor. One bold way to improve targeting effectiveness is to enforce means testing or targeting through indicators. A drawback is that it would interfere with the self-selection mechanisms that many believe is necessary for this type of intervention to work. A softer approach would be to change the way the model works to better fit the poorest. Implementers could ensure that very low savings amounts are possible in all groups. Or they could supplement VSLA intervention with other types of activities aimed at including the poorest, for example, the mimicking the graduation programs used by e.g. BRAC to enable the poorest to join regular microfinance (Halder and Mosley 2004, Bandiera *et al.* 2011, Hashemi and de Montesquiou 2011).

In discussions of microfinance there is a need to be more specific whenever we talk about reaching the poorest, as well as when we claim that microfinance is not for the poorest. If we want to reach a large fraction of the poor in rural Malawi, microfinance institutions are not a good idea. Community-managed microfinance is. But if we want to reach more of the poor than of the non-poor, then VSLA in its present form is not the right intervention, at least not in northern Malawi.

References

- Alatas, V., Banerjee, A., Hanna, R., Olken, B.A. and Tobias, J. 2012. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4), pp. 1206-1240.
- Alkire, S. and Foster, J. 2011. Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7), pp. 476-487.
- Allen, H. and Panetta, D. 2010. *Savings Groups: What Are They?* Washington DC: The SEEP Network.
- Allen, H. and Staehle, M. 2007. *Programme Guide - Field Operations Manual version 3.1*. Stuttgart: VSL Associates, available from <http://vsla.net> [Accessed October 2012].
- Amin, S., Rai, A.S. and Topa, G. 2003. Does microcredit reach the poor and vulnerable? Evidence from northern Bangladesh. *Journal of Development Economics*, 70(1), pp. 59-82.
- Armendariz de Aghion, B. and Morduch, J. 2005. *The economics of microfinance*. Cambridge, Mass.: MIT Press.
- Bandiera, O., Burgess, R., Das, N.C., Gulesci, S., Rasul, I., Shams, R. and Sulaiman, M. 2011. Asset Transfer Programme for the Ultra Poor: A Randomized Control Trial Evaluation *CFPR Working Paper*, No. 22.
- Barinaga, M. 1992. Profile of a field: neuroscience. *Science*, 255(5050), pp. 1366-1367.
- Bouman, F. 1995. Rotating and accumulating savings and credit associations: A development perspective. *World Development*, 23(3), pp. 371-384.
- Bouman, F.J.A. 1984. Informal Savings and Credit Arrangements in Developing Countries: Observations from Sri Lanka. In D. Adams, D. Graham and J.D. Von Pischke (eds.) *Undermining Rural Development with Cheap Credit*. Boulder, CO, USA: Westview Press, pp. 232-247.

- Chen, S. and Ravallion, M. 2010. The developing world is poorer than we thought, but no less successful in the fight against poverty. *The Quarterly Journal of Economics*, 125(4), pp. 1577-1625.
- Christen, R.P. 2001. Commercialization and mission drift - the transformation of microfinance in Latin America. Washington: Occasional Paper No 5, CGAP.
- Coady, D., Grosh, M. and Hoddinott, J. 2004. Targeting outcomes redux. *The World Bank Research Observer*, 19(1), pp. 61-85.
- Coady, D., Martinelli, C. and Parker, S.W. 2013. Information and participation in social programs. *The World Bank Economic Review*, 27(1), pp. 149-170.
- Conning, J. and Kevane, M. 2002. Community-based targeting mechanisms for social safety nets: A critical review. *World Development*, 30(3), pp. 375-394.
- Cornia, G.A. and Stewart, F. 1993. Two errors of targeting. *Journal of International Development*, 5(5), pp. 459-496.
- Cull, R., Morduch, J. and Demirgüç-Kunt, A. 2007. Financial performance and outreach: a global analysis of leading microbanks. *The Economic Journal*, 117(517), pp. F107-F133.
- Deaton, A. 1997. The analysis of household surveys: a microeconomic approach to development policy. Baltimore: Johns Hopkins University Press.
- Deaton, A. and Dupriez, O. 2011. Purchasing power parity exchange rates for the global poor. *American Economic Journal: Applied Economics*, 3(2), pp. 137-166.
- Foster, J., Greer, J. and Thorbecke, E. 1984. A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, 52(3), pp. 761-766.
- Foster, J., Greer, J. and Thorbecke, E. 2010. The Foster–Greer–Thorbecke (FGT) poverty measures: 25 years later. *Journal of Economic Inequality*, 8(4), pp. 491-524.

- Ghatak, M. and Guinnane, T.W. 1999. The economics of lending with joint liability: theory and practice. *Journal of Development Economics*, 60(1), pp. 195–228.
- Glewwe, P. and Jacoby, H. 2000. Recommendations for Collecting Panel Data. In M. Grosh and P. Glewwe (eds.) *Designing household survey questionnaires for developing countries : lessons from 15 years of the Living Standards Measurement Study*. Washington DC: World Bank, pp. 275-314.
- Grosh, M.E. 1994. Administering targeted social programs in Latin America: From platitudes to practice. Washington D.C.: World Bank.
- Halder, S. and Mosley, P. 2004. Working with the ultra-poor: learning from BRAC experiences. *Journal of International Development*, 16(3), pp. 387-406.
- Handa, S., Huang, C., Hypher, N., Teixeira, C., Soares, F.V. and Davis, B. 2012. Targeting effectiveness of social cash transfer programmes in three African countries. *Journal of development effectiveness*, 4(1), pp. 78-108.
- Hartarska, V. and Nadolnyak, D. 2007. Do regulated microfinance institutions achieve better sustainability and outreach? Cross-country evidence. *Applied Economics*, 39(10), pp. 1207-1222.
- Hashemi, S.M. and de Montesquiou, A. 2011. *Reaching the Poorest: Lessons from the Graduation Model*. CGAP Focus Note, no. 69. Washington D.C.: CGAP.
- Heckman, J.J. and Smith, J.A. 2004. The Determinants of Participation in a Social Program: Evidence from a Prototypical Job Training Program. *Journal of Labor Economics*, 22(2), pp. 243-298.
- Hermes, N., Lensink, R. and Meesters, A. 2011. Outreach and Efficiency of Microfinance Institutions. *World Development*, 39(6), pp. 938-948.
- Houssou, N. and Zeller, M. 2011. To target or not to target? The costs, benefits, and impacts of indicator-based targeting. *Food Policy*, 36(5), pp. 627-637.

- Hulme, D. 2000. Is microdebt good for poor people? A note on the dark side of microfinance. *Small Enterprise Development*, 11(1), pp. 26-28.
- IRIS Center 2012. Poverty Assessment Tool Accuracy Submission USAID/IRIS Tool for Malawi. Maryland: University of Maryland.
- Kennedy, P.E. 1981. Estimation with correctly interpreted dummy variables in semilogarithmic equations. *American Economic Review*, 71(4).
- Lang, C., Barrett, C.B. and Naschold, F. 2013. Targeting Maps: An Asset-Based Approach to Geographic Targeting. *World Development*, 41(1), pp. 232-244.
- Maes, J.P. and Reed, L.R. 2012. State of the microcredit summit campaign report 2012. *Microcredit Summit Campaign: Washington, DC*.
- Mohammed, R., Norris, A.H., Evans Alayne, M. and Timothy, G. 1999. Demystifying nonparticipation in microcredit: a population-based analysis. *World Development*, 27(2), pp. 419-430.
- Morduch, J. 1995. Income Smoothing and Consumption Smoothing. *The Journal of Economic Perspectives*, 9(3), pp. 103-114.
- Navajas, S., Schreiner, M., Meyer, R., Gonzalez-Vega, C. and Rodriguez-Meza, J. 2000. Microcredit and the Poorest of the Poor: Theory and Evidence from Bolivia. *World Development*, 28(2), pp. 333-346.
- Niehaus, P., Atanassova, A., Bertrand, M. and Mullainathan, S. 2013. Targeting with Agents. *American Economic Journal: Economic Policy*, 5(1), pp. 206-238.
- Park, A., Wang, S. and Wu, G. 2002. Regional poverty targeting in China. *Journal of Public Economics*, 86(1), pp. 123-153.
- Ravallion, M. 2001. Growth, inequality and poverty: Looking beyond averages. *World Development*, 29(11), pp. 1803-1815.

- Ravallion, M. 2009. How Relevant Is Targeting to the Success of an Antipoverty Program? *World Bank Research Observer*, 24(2), pp. 205-231.
- Ravallion, M. 2011. On multidimensional indices of poverty. *The Journal of Economic Inequality*, 9(2), pp. 235-248.
- Ravallion, M. 2012. Why Don't We See Poverty Convergence? *American Economic Review*, 102(1), pp. 504-523.
- Rutherford, S. 2009. *The Poor and Their Money*, 2nd ed. London: Practical Action Publishing.
- Sen, A. 1976. Poverty: an ordinal approach to measurement. *Econometrica: Journal of the Econometric Society*, pp. 219-231.
- Soe, L. and Elaine, K.Y. 2008. What's wrong with the pipeline? Assumptions about gender and culture in IT work. *Women's Studies*, 37(3), pp. 176-201.
- Solon, G., Haider, S.J. and Wooldridge, J. 2013. What Are We Weighting For? *NBER Working Paper*.
- UNDP 2010. Technical note 1. Calculating the Human Development Index. In Human Development Report 2010 - 20th Anniversary Edition. The Real Wealth of Nations: Pathways to Human Development. New York: UNDP.
- White, K. 2004. The leaking pipeline: Women postgraduate and early career researchers in Australia. *Tertiary Education & Management*, 10(3), pp. 227-241.

Chapter 3

Small groups, large profits:
calculating interest rates in
community-managed
microfinance

Small groups, large profits: Calculating interest rates in community-managed microfinance

Ole Dahl Rasmussen

University of Southern Denmark and DanChurchAid

Abstract: Savings groups are a widely used strategy for women's economic resilience – over 80 per cent of members worldwide are women, and in the case described here, 72.5 per cent. In these savings groups it is common to see the interest rate on savings reported as '20-30 per cent annually'. Using panel data from 204 groups in Malawi, I show that the correct figure is likely to be at least twice as much. For these groups, the annual return is 62 per cent. The difference comes from sector-wide application of non-standard interest rate calculations and unrealistic assumptions about the savings profile in the groups. As a result, it is impossible to compare returns in savings groups with returns elsewhere. Moreover, the interest on savings cannot be compared with the interest rate on loans. I argue for the use of a standardised comparable metric and suggest easy ways to implement it. Development of new tools and standards along these lines are fortunately under way from key players in the sector and should be welcomed by donors, politicians and practitioners to improve transparency and monitoring.

Keywords: microfinance, interest rates, performance monitoring, community-managed microfinance, village savings and loan associations, Malawi, NGO

Acknowledgements: The author would like to thank CARE Malawi for providing the data, Hugh Allen for continuous feedback, Nikolaj Malchow-Møller for comments on an earlier draft, Knud Rasmussen for helpful insights and two anonymous referees for useful suggestions.

This chapter has been published as a journal article. The full reference is: Rasmussen, O.D. 2012. Small groups, large profits: Calculating interest rates in community-managed microfinance. *Enterprise Development and Microfinance*, 23(4), pp. 298-318.

Introduction

In 2011 the weekly newspaper *The Economist* called savings groups ‘the hottest trend in microfinance’ and wrote: ‘...returns on savings are extremely high – generally 20-30 per cent a year. Borrowers typically pay interest rates of 5-10 per cent a month’ (The Economist Dec 2011). In this paper I demonstrate how savings groups enable some of the poorest Africans to earn not 30 per cent but 60 per cent interest on their savings. I do this by applying the most widely used financial calculation of interest rates to panel data on 204 savings groups with 3,544 members in Malawi, finding the median interest on savings to be 62 per cent per year, or 3.8 per cent per month. This figure is directly comparable to the 10 per cent monthly interest rate on loans.

The discussion is important because the current calculation makes interest rates on savings incomparable to the interest rate on loans as well as to interest rates elsewhere. De Mel *et al.* (2008) have shown that returns in Sri Lankan microenterprises are 60 per cent. If it were true that savings groups generate only 30 per cent, Sri Lankan savers should keep their money in microenterprises and not join savings groups. But in finding that savings groups generate returns of 60 per cent, the case is different. De Mel *et al.* do not provide information on savings alternatives in Sri Lanka, but the imagined example illustrates why the difference matters. It is no coincidence that transparent pricing is at the core of current efforts to ensure client protection in microfinance: people should know what they pay, and also what they get. Finally, non-standard interest rate calculations make monitoring difficult: when loans and savings are not comparable, how do we know the good groups from the bad ones?

Savings groups have at least 1.5 million members worldwide, primarily in Africa. Since more than 80 per cent of members worldwide are women (72.5 per cent in the case described here), the topic is particularly relevant to discussions on women’s economic resilience. If the savings groups reporting on the portal <http://savingsgroups.com> were a single microfinance institution, this institution would be the ninth largest in the world and the second largest in Africa with regard to total number of savers.

In the next two sections, I discuss a relevant argument against the standard financial calculation of interest -- the fact that interest rates in general and compounded interest in particular are alien to most cultures where savings groups thrive. After a general discussion of interest rate calculations, I turn to data and analyse the interest rate in 204 Malawian savings groups, each observed four times during one year. I also develop a look-up table that enables use of the standard financial calculation with very little computation. After this, I provide an example of how the new interest rate metric can be used to monitor groups. Since the proposed calculation enables direct comparison of interest rates on loans and on savings, the difference between these two can be used to spot groups in trouble, something project managers can use to direct attention to the right places.

Apart from the positive message that interest rates on savings are twice as high as we thought, my primary recommendation is to acknowledge the advantages of the standard financial calculation as described in Annex 1. In situations where we cannot use this calculation, we should at least use the best possible approximation. Fortunately, the leading provider of monitoring tools for savings groups, VSL Associates, has decided to change its calculations, taking some, though not all, of these issues into account in the tools currently being developed. As such, the present paper serves to justify why this change is needed and why it should be adopted by the rest of the savings group sector.

Are savings groups different?

Savings groups are groups of 15 to 25 people, the majority women, who are taught how to manage their own funds. Their precise way of working is carefully described elsewhere (Allen and Panetta 2010), but typically savings are made on a regular basis (e.g. weekly) and the saved capital is lent to the group members. Loan duration is commonly three months and all the groups' assets are shared out once every year or so according to the individual level of savings. These groups are usually not regulated, and people working with savings groups commonly refer to them as being 'under the radar' of national

supervision. Some perceive this as an advantage, whereas others point out that groups under the radar can avoid legislation which requires them to provide about information to consumers, in particular the interest rates on loans and savings (Rhyne and Rippey 2011). Partly because of this, interest rate calculations have not followed the standards and practices used elsewhere. In the next section I will discuss why I think that standard financial calculations should be followed, even for savings groups.

To compound or not to compound

The history of interest is long: Greece in 600 BC and Babylon 1,200 years earlier had interest rates and legislations regulating them. Nevertheless, research in anthropology suggests that the concept of interest rates originated in the countries that are today donors of foreign aid (Homer and Sylla 1996). As a result, the concept might be alien to most savings group members. Certainly, compounded interest is an intrinsic part of the culture of finance in many high GDP countries, so much so that the method I propose below is not just used in most textbooks on finance, but has also been made into law in many jurisdictions, including the EU and USA (EU Directive 2008/48/EC 1998, Truth in Savings Act 1991, Brown and Zima 2011, MFTransparency 2011).

What is so special about compounded interest? It multiplies at an ever-increasing rate. If I invest 100 shillings for two months and get 225 shillings at the end of the period, then I have earned interest of 125 per cent in the two-month period. The monthly interest, however, is 50 per cent, not 62.5 per cent, since my investment is comparable to one with a 50 per cent monthly return during two months. Methods for compounding use concepts of net present value and internal rate of return as described in Annex 1.

David Graeber has shown how, historically, all interest was considered usury and has argued that promoting interest rate-bearing loans might lead to social problems (Graeber 2011). In contrast, other anthropological studies of indigenous rotating savings and credit associations have found that interest rates are a natural part of the way

these groups work (Ardener and Burman 1995). While the discussion for and against interest rates as such is relevant to microfinance as a whole, I will not go further into the discussion here. Interest rates and the reporting of interest rates are a part of current savings group practice. What I will discuss is how to report interest rates.

To frame the discussion, it is useful to ask Robert Chambers' rhetorical question: whose reality counts? Clearly, participants' reality ultimately counts and therefore local understanding of interest is relevant. Shipton (2010) notes that few African languages have words for interest or usury and thus the very concept often lacks an indigenous counterpart. Where there are concepts similar to interest, they relate to ratios, not to rates. Indeed this is probably why flat-rate interest is so common in microfinance: it is locally accepted.

Any development intervention is an ongoing negotiation of concepts and the very implementation of savings groups is an example of this negotiation. It resonates with indigenous financial arrangements, while at the same time explicitly aiming to augment these. No matter what, savings groups will change local cultures. The way interest rates are computed is a part of this change and it should be clear how and why they are calculated the way they are. Neither international standards nor local customs are in and of themselves good reasons for adopting one calculation over the other. Here the key advantage of the financial method is the ability to compare monthly and annual rates as well as rates of different providers. Few savers in Europe would be able to re-calculate the interest they get on their savings account, since it is done exactly following the methods in Annex 1. Only 18 per cent of adult Britons answered correctly when asked a simple question on compounded interest (Lusardi and Mitchell 2007). They rely on legislation and public oversight to ensure correct calculations. It seems illogical that the absence of these mechanisms should be a reason for computing and communicating incomparable interest rates, if comparison is indeed needed. In the following, I will expand on the reasons why it might be useful.

In semi-mature microfinance markets such as Cambodia or Kenya, savings groups exist alongside conventional finance. Here a direct

benefit is that participants are able to compare the services of the savings group with those of ordinary providers. There are many reasons for members to choose their savings groups, but Collins *et al.* (2009) have documented how interest rates are at least one of the reasons. The ability to compare savings groups with other providers is one direct benefit to members of the financial calculation suggested here. A potential caveat is that there is homogeneity in savings and borrowing patterns inside the group. However, in a survey of 1,775 households in Northern Malawi, 54 per cent of group members indicated that they had not taken a loan during the past year, despite being members of a savings group throughout this period (Ksoll *et al.* 2013, author's calculations). This points to heterogeneous intra-group savings and borrowing practices.

There are also indirect benefits. Project managers need to assess performance of the group to support the groups in need. One performance metric is money not accounted for in the groups. If savings are lent out at ten per cent per month, savers should in principle be able to take away ten per cent per month – that is, in principle, since many other things happen along the way. Taking this into account, comparable interest rates will allow supervisors to separate credit and savings from loss. This can only be done if the metrics are *internally consistent*; that is, when interest rates on savings and interest rates on loans have the same meaning within the reporting system.

Non-standard interest rate calculations

So what is wrong with the typical interest rate calculations? There are two issues: first, the interest rate calculation itself; and, second, annualisation of this rate whenever groups are less than a year old. I go through these calculations below.

Interest rate calculation

The calculation used for annual interest rates in the current management information system developed by CARE, Oxfam, and CRS as well as on <http://savingsgroups.com> is this:

$$\text{Returns on savings} = \frac{\text{Net profit/loss}}{\text{Cumulative value of savings}} \quad (1)$$

In this paper, I call this the simple method. To illustrate the pitfalls of this method, let's look at two groups, assume that we know their savings profile, i.e. when they save. Both save 1,000 shilling and end up with 1,100 after one full year. In Group A, everyone saves everything on 1 January. The members have to live without their 1,000 shillings for the entire year. Group B postpones saving until 1 December, at which time it saves 1,000 shillings. The members of Group B must get by without their 1,000 shillings for one month. On 31 December the same year, both groups have total assets of 1,100 shillings and profits of 10 per cent of their cumulative savings. The groups' savings profiles are illustrated in Figures 1 and 2.

Calculating interest using equation (1) above, both groups yield an interest rate on savings of 10 per cent per year. As should be clear from the two examples, however, the interest rate on savings in the groups is not the same. In Group A, members get 100 shillings in profits when they save their money for a year; and in Group B, they get the same in just one month. Clearly, Group B yields a higher interest rate. Following financial interest calculations, the annual real interest rate would be 10 per cent for Group A, but a staggering 214 per cent for Group B.

The central point is that the savings profile matters greatly to the interest rate when we want to use the generated profits as a basis for calculation. To calculate the interest rate, we must assume a savings profile. For the present purpose, this raises two questions: what are the assumptions about the savings profile in equation (1) above, and what might the real savings profile be in savings groups?

Figure 1. Group A saves everything in the beginning

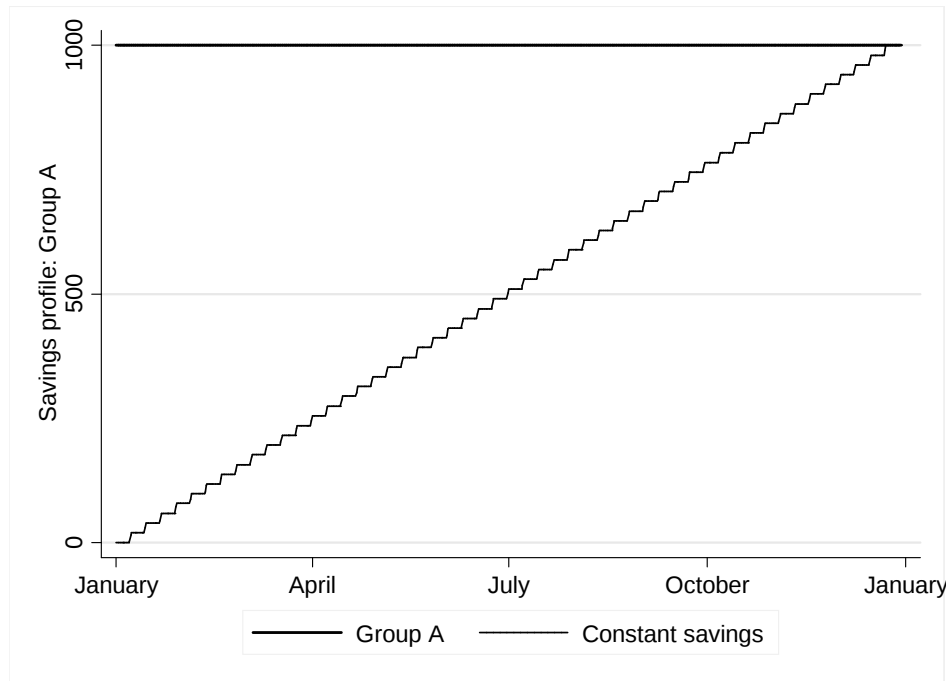
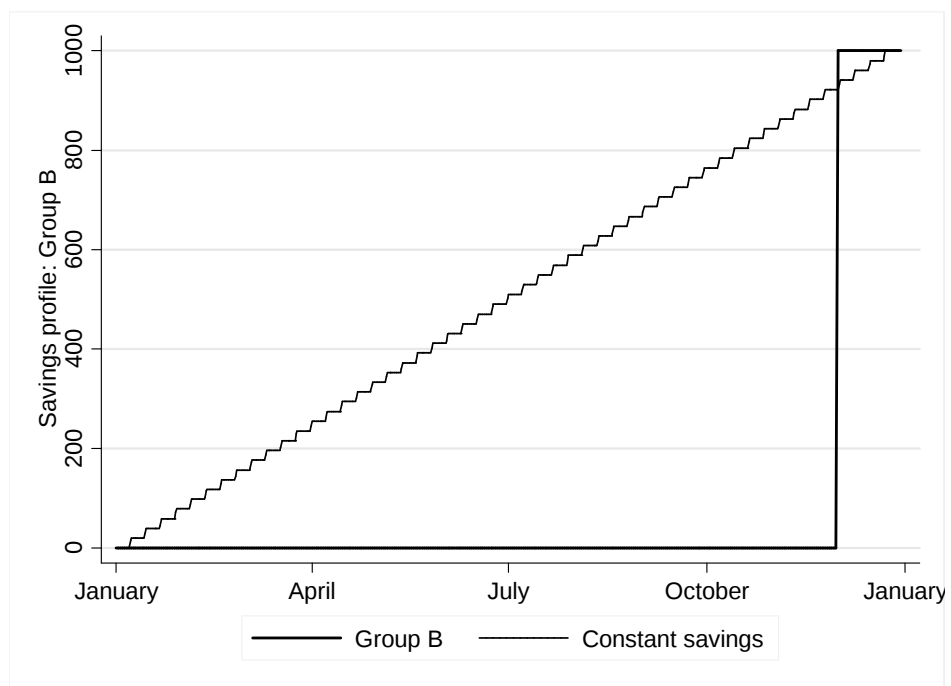


Figure 2. Group B saves everything on December 1st.



The assumption underlying the simple method is that the savings profile is exactly as in Group A. The only case where the formula is correct is when everything is saved at the beginning of the year and kept in (and lent out by) the group until payout at the end. Turning to the second question, Group A's savings profile is unrealistic for savings groups simply because of the way the groups work. Savings are carried out by purchasing so-called shares in the groups, and members are strongly encouraged to buy at least one share, but internal rules prohibit them from buying more than five shares per week. Nevertheless, the exact savings profile is an empirical question which is analysed in the next section.

Annualisation

The second non-standard calculation is annualisation. The management information system uses the following formula for annualising the interest rate:

$$\text{Annual interest rate} = r_{\text{period}} * \frac{52}{n} \quad (2)$$

where r_{period} is the period interest rate and n is the length of the period in weeks. This formula ignores compounded interest. Savings groups require payment of interest every month and even if only interest is paid, this corresponds to compounding. If interest rates are to be externally and internally consistent as discussed above, then compounded interest must be taken into account. The standard formula for calculating the annualised interest from the available data would be to first calculate the weekly interest and then annualise. The formulas are:

$$r_{\text{week}} = (r_{\text{period}} + 1)^{1/n} - 1 \quad (3)$$

$$r_{\text{annual}} = (r_{\text{week}} + 1)^{52} - 1 \quad (4)$$

Table 8 illustrates this for 13 periods, which is the number of four week periods in a year. The initial loan amount is 100 shillings and monthly interest is 10 per cent. Each new period's loan amount is simple the preceding period's loan amount plus interest. The annual interest is 245 per cent, since a savings group year has 13 four-week periods.

Table 8. Compounding interest

	Loan amount	Interest
Period 1	100	10
Period 2	110	11
Period 3	121	12
Period 4	133	13
Period 5	146	15
Period 6	161	16
Period 7	177	18
Period 8	195	19
Period 9	214	21
Period 10	236	24
Period 11	259	26
Period 12	285	29
Period 13	314	31
Repayment	345	
Total interest	245	

Data on interest rates and savings profiles

As mentioned in the previous section, the savings profile is essential when calculating the interest rate. Certainly, the savings profile assumed by the formula currently used to compute interest rates is unrealistic. But what, then, is realistic? To answer that question, we must turn to data. The data I use here are collected every three months and contain information on, for example, total savings, total assets and group age. Because of the multiple time points, the data give an indication of the savings profile. More time points would give more precise information. In effect, the data tell us about the overall interest on savings in the groups. After an overall description of the data, I compute the real interest rate using the standard financial calculation.

Then, I primarily compare these calculations with three approximations: the current simple method, a calculation assuming constant weekly savings and a variation of the simple methods which is currently being rolled out. I call the latter the new simple method.

The data

The initial data encompass 974 groups. The groups are observed every three months and most groups have more than one observation. To provide information on the savings profile, I need multiple data points, so I exclude the groups with less than four observations. That leaves 239 groups that have been observed during the course of at least a year. Following common practice in savings groups, the groups distribute all funds once per year at the end of a so-called cycle. After one cycle, it is common to start again with a larger initial savings contribution from all members. Since I do not have precise data on the funds involved in the distribution, I limit the analysis to one such cycle, lacking information about any initial savings, I use only first-cycle groups. This leaves 204 groups for the analysis. One concern in this trimming exercise is that the subset of 204 groups is not representative of the larger pool of groups. Well-run groups might have different savings patterns to poorly managed groups. The well-run groups might also live longer and thus have a higher probability of entering my analysis. Because of this, the results are valid only for groups that survive more than one year. In practice, groups are usually considered independent after operating for one year, so the results can be thought of as valid for independent groups. Groups elsewhere might be different from the groups studied here, in which case the interest rate might also be different in other contexts. Comparing the simple interest rates with savingsgroups.com shows that these groups are similar to the global average in that respect.

Before I turn to the analysis of the real savings profile, it is worth mentioning some descriptive statistics of the groups. These are listed in table 9. For example, members buy an average of 82 shares in a period of 40 weeks, or two shares per meeting. The groups have been trained partly by field officers paid for by the project, and partly by

village agents paid for by the groups themselves. The average number of loans per group during the whole cycle is 7.9, and 9.1 members or 53 per cent in each group are savers only at the time of data collection.

Table 9. Descriptive statistics

	Total number of groups	Total number of observations	Mean	SD overall	SD between groups**	SD within groups**
Group size at start of cycle	206		17.7	3.00		
Group size at the last data collection	153		17.10	3.50		
Average savings per member (Malawi Kwacha)	206		6367	3671		
Average profits per member (Malawi Kwacha)	206		1801	1522		
Share of women in groups	206	787	72.5%	23.7%	22.6%	6.3%
Dropouts since start of cycle	153		1.13	2.20		
Number of outstanding loans per group	206	840	7.9	6.8	3.0	6.1

* The number is sometimes lower than 206/840 due to missing observations in the original data

** The standard deviation between groups is the standard deviation of the group averages from the overall average. A high value means a big difference from one group to another. The standard deviation within groups is the standard deviation of the individual group's four observations from the group's average. The two figures are only defined for measures that change over time.

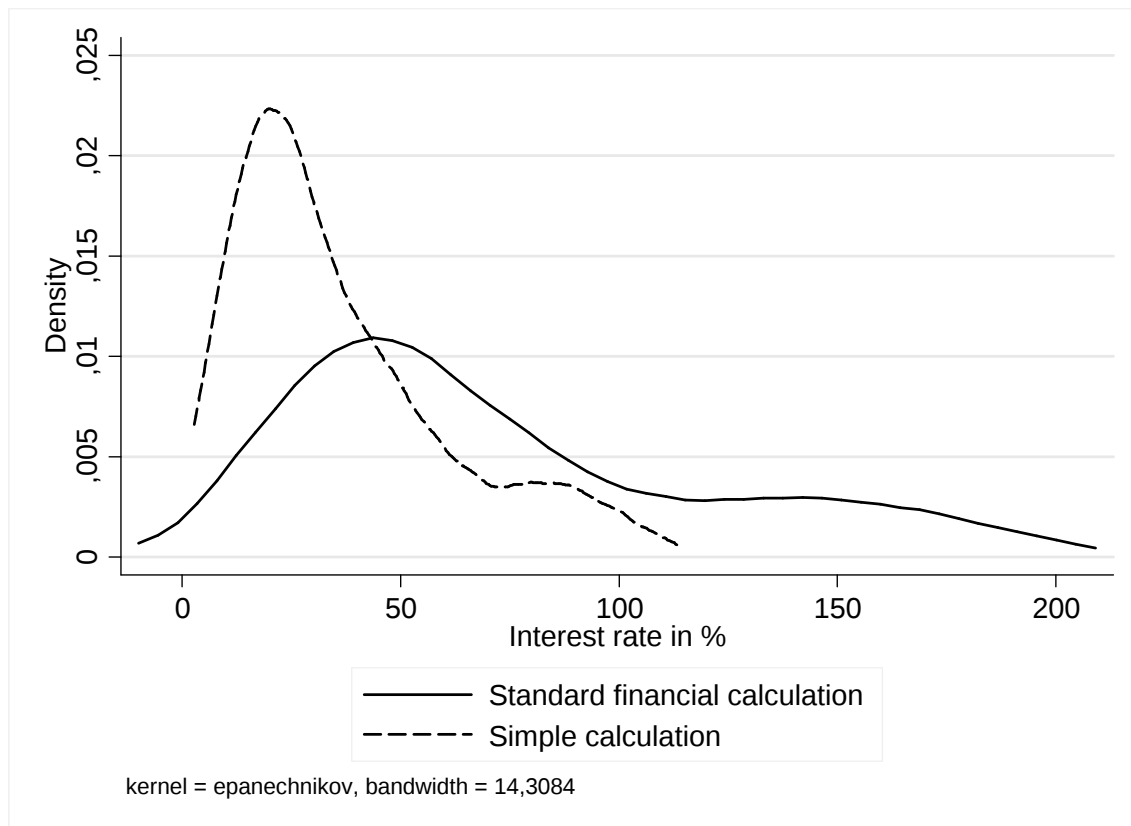
The median real interest rate is 62 per cent

The key result is the effective interest rate calculated using standard financial calculations based on net present value and internal rates of return as explained in Annex 1. In practice, it takes into account the savings profile and the problems of the simple method mentioned in the example with Group A and Group B. Using this method, I calculate one effective interest rate per group using all four observations from each group. I assume that the rate of savings is constant between the dates on which the observations were made, but since there are four observations, the overall savings profile can be non-linear. If groups save more in the beginning and less at the end,

this is taken into account. As such, this way of computing the interest rate imposes a minimum of assumptions on the savings profile given the available data. Of course, if the individual savings profiles differ from the group level average, the individual interest rates will be different.

Using this method, the median annual real interest rate in the 204 groups is 62 per cent, or 3.8 per cent per four-week period. In contrast, the simple method generates a median interest rate of 29 per cent. The averages are 88 per cent for the financial method and 36 per cent for the simple method, but since there are a few high interest rates and many rates between zero and one, I consider the median to be a better descriptive measure in this case. Taking the financial method as the standard, the simple method is wrong by at least 18 percentage points in more than 75 per cent of the cases. It is within plus or minus 10 percentage points in only 8.3 per cent of the cases. Figure 3 displays the density plots of the two distributions. It is clear that the simple calculation falls systematically below the financial calculation. (In Figures 3 and 4, I remove between one and 16 outliers for the purpose of display. These are, however, included in the calculations.) This is a result of the fact that savings do not happen in the beginning of the cycle as the simple method assumes, but throughout the period. As such, the simple calculation underestimates the true interest rate. Further, the simple interest rate is less variable than the standard financial calculation. I conclude that the approximation used by the simple method misses the mark.

Figure 3. Density plot of the two ways of computing interest rates.



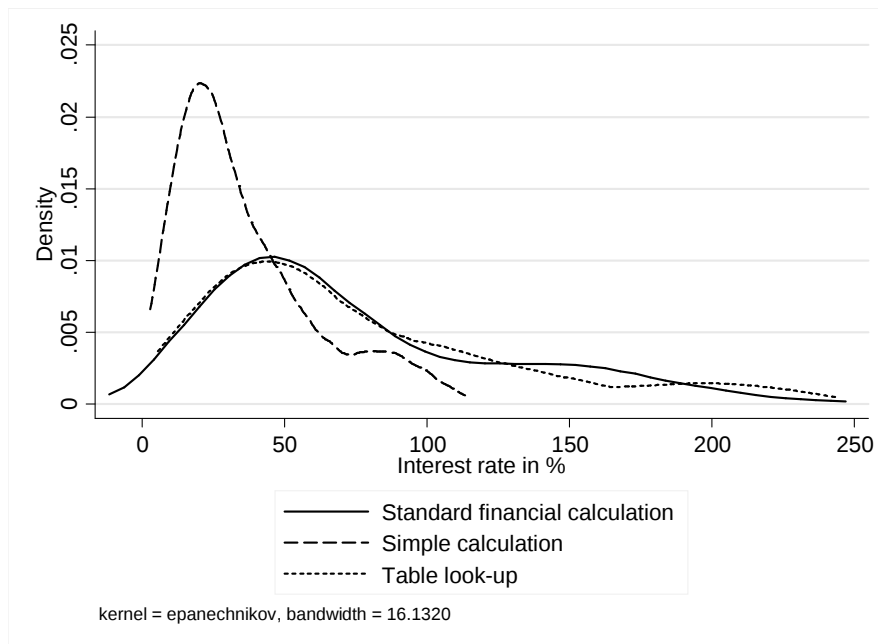
How should we report interest rates on savings in savings groups?

The standard financial calculation requires computations that might seem difficult to implement as part of everyday savings group practice. Therefore, it is not practical for groups to change the way they calculate their own returns during share-out, since this process must be manageable by the group themselves and it is complicated by the fact that information from more than one time point is needed. To move forward, two options seem feasible. First, the next generation of the standard management information system is currently being developed and will use data from multiple observations per group, making direct financial calculation possible. Alternatively, one can assume constant savings and then compute the interest rate from one observation using the age of the group and the profits. In this case, the calculation can be done once and displayed in a table. This table can

then be used for look-up of the real effective interest rate. Such a table is presented in Annex 2 for annual savings and Annex 3 for monthly savings. The latter is included since the interest rate will inevitably change during the cycle, and an annual rate might give the false impression that it is a projection. I use numbers from the table with annual interest rates. Compared with the standard financial calculation, the table is imprecise in two ways. Most importantly, it ignores the fact that some groups save more in the beginning of the cycle, while others save more at the end. This changes the effective interest rates. Moreover, the table must necessarily include discrete steps, in this case of 2.5 percentage points and four-week periods.

To investigate the feasibility of the look-up table, I used it to arrive at the interest rates for all 204 groups and compared them with the financial calculation. In general terms, the look-up table is a much better approximation than the simple calculation. Where the simple calculation falls far below the financial calculation, the look-up gives interest rates both below and above the financial calculation. As such, the median interest rate obtained using the table is 58 per cent, compared to 62 per cent using the standard financial calculation. The average is 85 per cent compared to 88 per cent. The difference between the look-up measure and the standard financial calculation falls around zero. In contrast, the difference between the simple and the financial calculation is systematically below zero. In fact, the error is 3.1 percentage points on the average, and the median of the error is 1.3 percentage points, a figure not statistically significant from zero at a 5 per cent level. For half of the 204 groups, the difference is between 1.1 and 3.7 percentage points. The density plots of all three calculations are shown in Figure 4.

Figure 4. Density plot of the three ways of computing interest rates.



The reason for the match between the look-up table and the standard financial calculation is that, on average, the savings profiles are linear. To compare savings profiles, I have normalised both age and savings so that both are between zero and 100. This enables graphing of the savings profile. Five examples are given in Figure 5 and all 200 groups are graphed in Figure 6 to provide an illustration of the comparison of medians above: the fact that linear or constant savings is a good average approximation.

Figure 5. Five examples of savings profiles

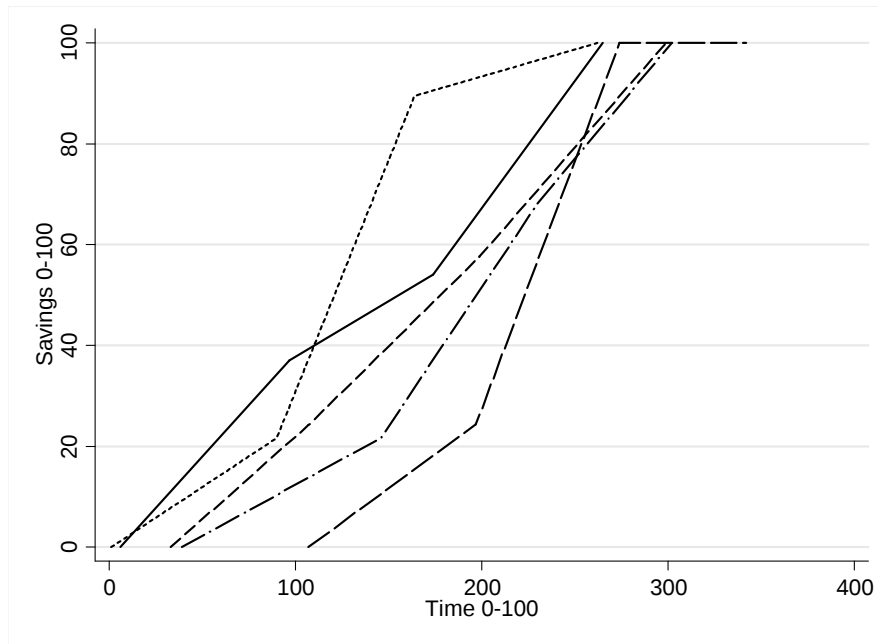
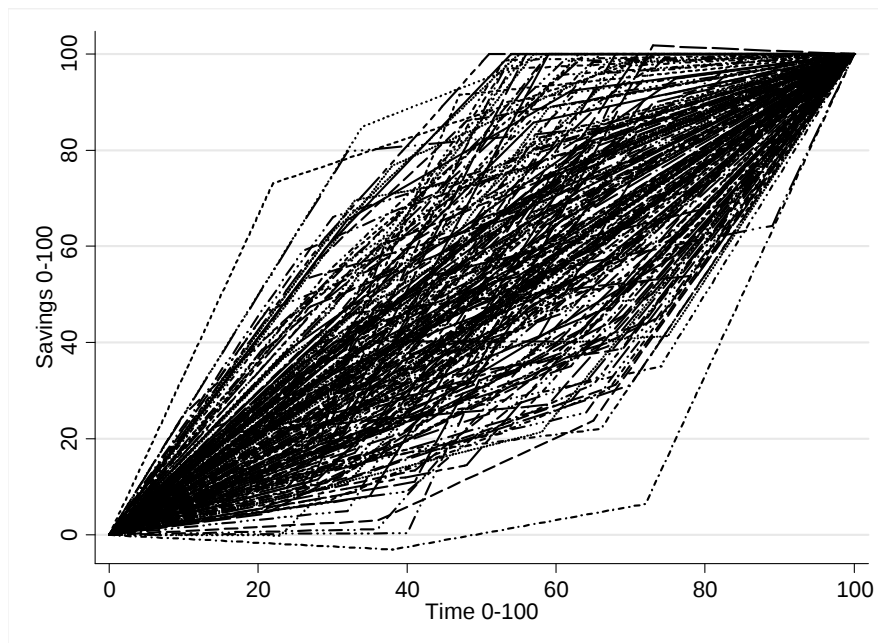


Figure 6. All 204 savings profiles.



A new simple method

In the monitoring and information system that VSL Associates is planning to launch in 2013, returns will be calculated using this metric:

$$\text{Returns on average savings} = \frac{\text{Net profit}}{\text{Cumulative value of savings}/2} \quad (3)$$

The key difference from Equation (1) is that cumulative savings is divided by two. It is easy to see that this will simply double the returns figure and as such, make it much closer to the financial calculation. That there are still important differences can be seen by using the examples from Group A and Group B above: even with the new figure, these two groups will have the same interest rate, because the savings profile is not taken into account. Averaging out savings assumes that early and late savings count the same. In the financial calculation, they do not and therefore, the new simple method is systematically biased downwards. On average, in the 204 Malawian groups, the new simple calculation falls 16.2 percentage points below the financial calculation.

Using comparable interest rates for monitoring

Even with the adjustment to the interest rate on savings obtained by the financial calculation above, there is still an important unexplained fact. The annualised interest rate on loans is 245 per cent, but the annualised interest rate on savings is 62 per cent. Where does this large difference come from? An obvious explanation is incomplete fund usage, which I analyse in the next section.

Incomplete fund usage

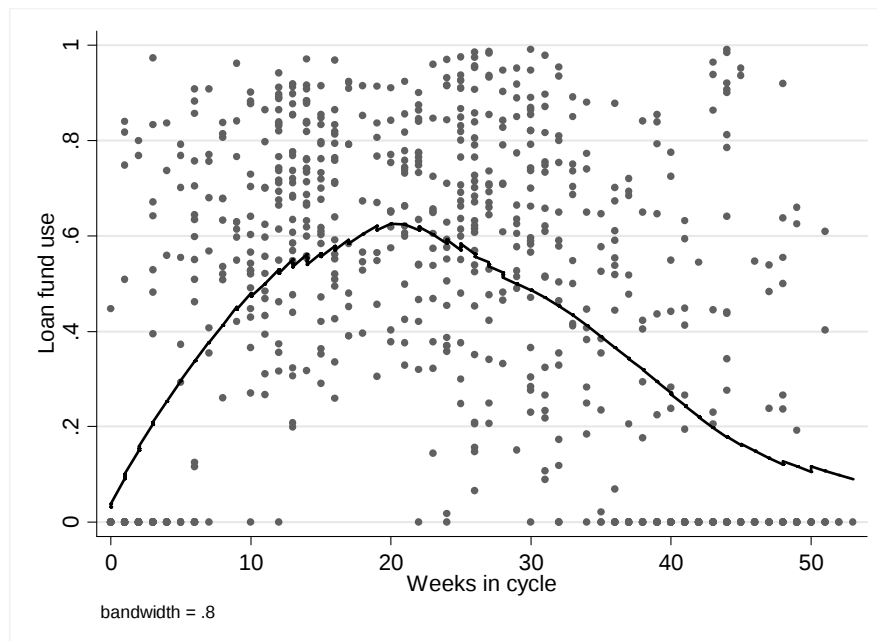
When members save in their groups, not all savings are lent out at any one time. Naturally, the funds not lent out do not accumulate interest even though they are saved. The ratio between all funds and lent out funds is known as the loan fund use ratio. So the expected interest rate is not 245 per cent, but lower. How much lower? The interest rate on

loans is set to 10 per cent or 20 per cent with a few groups higher or lower. The average is 14.9 per cent or 511 per cent annually. The average loan fund use is 46 per cent across all observations and all groups. A simple adjustment gives an annual interest rate of 230 per cent, but this figure depends on how the loan fund use ratio develops over time. The change in loan fund use is shown in Figure 3. The solid line shows the local average of loan fund use. Clearly, it is low in the beginning and at the end. At the same time, the nominal interest rate is normally constant and the savings level increases. To enable performance monitoring, I combine these three pieces of information in a ‘potential interest rate’ for each group; that is the return on savings the groups would have made if their stated nominal interest rates and loan fund uses are correct. To arrive at this figure, I compute the loan fund used at each loan meeting by assuming that loan-fund use develops linearly between the observation points. Using total savings, I compute how potential assets would accumulate. Using the same method as described in Annex 1, I then replace the final payout in the sequence of payments with the potential assets and re-compute the effective interest rate. This figure is the potential interest rate. Using this information, the average potential interest rate is 109 per cent, which is considerably lower than the 230 per cent from above. Nevertheless, the calculation does not change the basic finding that some funds are missing, since there is still a gap of 47 percentage points between 62 per cent and 109 per cent. In fact, the difference between the effective interest rate and the potential interest rate, which I call *the missing interest rate*, is an excellent metric to track performance of the groups because it measures missing money in the groups. Project managers should visit the groups with a missing interest rate of more than, say, 20 percentage points. In the present dataset this means that 168 groups should be checked.

The global data looks somewhat similar. Here I have only the simply average loan fund use, which is 52.2 per cent. Assuming a ten per cent nominal interest rate per four-week period, the corresponding annual interest rate is 110 per cent, but taking changing loan fund use into account might reduce this. Moreover, the reported interest rate on

savings of 35 per cent might be twice as high. Owing to the lack of information, it is not possible to confirm that the missing interest rate is lower than in my sample, but on the other hand it cannot be ruled out.

Figure 3. Loan fund use over time (lowess plot)



Explaining this gap is not the purpose of the present paper and is likely to require more detailed information (for example from the passbooks used by many savings groups), but four possibilities seem particularly likely. First, there might be a large share of non-performing loans in the groups. Write-off of these loans is supposed to be reported, but for the 204 groups in my analysis, only one reported any write-off at all. It is possible that the groups have not yet given up on the loans, but they have not been paid back. Second, groups might have very relaxed repayment schedules without charging additional interest. Flexibility might very well be an advantage of savings groups, but this would lower the interest rate accordingly. Third, the surplus might simply have been stolen, and, finally, there might be issues of data quality. Which of these explanations is the correct one is

an area for future research and will require new data, for example from passbooks.

Conclusions

It is clear from the above treatment of interest rates in savings groups that these groups offer net savers very high interest rates, and are more beneficial to their members – primarily women – than has been previously reported. Looking at more than 200 groups in Malawi with a membership of 72.5 per cent women, I find a median interest rate of 62 per cent, and a mean of 88 per cent or monthly returns between 3.8 per cent and 5.0 per cent. Data from elsewhere in Malawi show that more than half might be net savers. The officially reported figures had a median of 29 per cent and a mean of 36 per cent. If net savers are among the poorest, this is very good news. For once, the most world's most marginalised and poor people are getting the world's best deals. I have argued that comparable metrics are relevant even if local approaches to calculating interest do not include compounding. On the contrary, the rapid growth of credit-led microfinance makes the possibility of comparison even more important for members as well as policy makers. Using the standard financial calculation is a prerequisite for comparison.

Acknowledging that current computing power in savings group implementation is probably limited, I offer an alternative approach to calculating all interest rates: a look-up table with translations from the simple interest rate in current use to the financial calculation. This method is considerably closer to the real financial calculation. On the basis of this interest rate, I suggest a metric for monitoring group performance, the missing interest rate, which compares the interest rate on loans with the interest rate on savings.

In practice, these results will matter to policy-makers, donors and practitioners in aid. First, practitioners should acknowledge that interest rates computed using standard financial calculations are the proper way to summarise the price or yield of money in time. As already mentioned, there are signs that a tool with improved metrics will be available in 2013. The present analysis is a justifies adopting

this revised tool, since it is superior to the old one. Adapting the standard financial calculation or the suggested look-up table would, however, be better still, and would enable practitioners to assess group performance by comparing interest on loans with interest on savings to obtain the 'missing interest'. Second, donors should sustain their funding for savings groups. Returns of 60 per cent should not be ignored. Even if the upcoming impact studies of savings groups show no effect, donors should fund studies of longer duration and greater statistical power using knowledge of how groups work gained from the first impact assessments. A high-yielding savings product needs serious assessment. Third, the comparison of interest rates on loans and savings enables calculation of the missing interest rate which in turn can be used to identify groups where money disappears. Everybody involved in savings groups should have an interest in finding out where missing funds go. Finally, an obvious question for future research is to look at factors that drive the effective interest rate.

After having been hidden for a long time, savings groups have started to appear on the radar of policy makers and academics. This is good news for everyone saving in, or working with, a savings group. But it will necessarily lead to comparison of savings groups and other financial instruments and will thus require practitioners to report key metrics, such as interest rates, in ways comparable to other products in the financial landscape. Only by doing that will savings groups appear on the radar as a viable alternative or cost effective supplement to formal financial access.

Annex 1. The standard financial calculation

How is the standard financial calculation different?

The financial calculation used computes effective interest using net present value, which is the sum of all payments discounted to the present. The effective interest rate is the interest rate where the net present value is zero. This method is used in financial textbooks as well as mandated by financial legislators around the world. It is also recommended by Microfinance Transparency, an organisation working for transparent interest rates. Annualisation is done using standard compounded interest.

For VSLAs, it makes sense to use weeks as the smallest period. Then the formula is:

$$NPV = \sum_{n=0}^N \frac{C_n}{(1+r)^n} = 0$$

where NPV is net present value, N is the total number of weeks and C_n is the payment in period n . For savings, this is a negative number. In period N , i.e. the period when the group is observed, this is a positive figure indicating how much the group could share out today. I find the weekly interest, r , by calculating a sequence of payments for each group, assuming that the sequence is linear between the individual observations and between the group start and the first observation. Since all groups are in their first cycle, they must have started with zero savings.

Calculation example

I use the values for assets and total savings. Assets are all savings and net accumulated earnings in the group. In the second and subsequent cycles, it is common for groups to contribute a large amount at the

first savings meeting, in which case the calculations below would not be valid. In my calculations, I have disregarded groups in their second cycle (this only concerned two groups). I calculate the savings per period and the average savings per week. The latter might vary over the total period. One unexplained fact for the group in the example below is the difference between assets and total savings in week one. I interpret this as a gift and thus do not count it as initial savings. The payout is total assets in the last period. A1 shows numbers from one group as an example.

Table A1. Example of information used to calculate the effective interest rate.

Weeks in cycle	Assets	Debt	Total savings	Savings difference	Week difference	Savings difference per week	Payout
1	16750		13000	13000			
14	104700		89600	76600	13	5892	
25	145000		124600	35000	11	3182	
38	278150		219500	94900	13	7300	278150

Simple calculation

$$\text{Interest rate (37 weeks)} = \frac{\text{Assets} - \text{Total savings}}{\text{Total savings}} = \frac{58650}{219500} = 26.7\%$$

$$\text{Interest rate (annualized)} = 26.7\% * \frac{52}{37} = 36.6\%$$

Financial calculation

From the above values, I make the sequence of payments as displayed in Table A2.

I then calculate the weekly interest rate using the NPV-formula above. This can be done in Excel using the functions ‘internal rate of return’ or ‘goal seek’. In my case, I use the statistical software Stata. All methods give the same results.

Interest rate (weekly) = 0.01191

Interest rate (37 weeks) = $(1+0.01191)^{37}-1=55.0\%$

Interest rate (annualized) = $(1+0.01191)^{52}-1=85.1\%$

Table A2. Calculation of effective interest rate setting net present value to zero.

Payment Discounted			Period			Payment Discounted			Period			Payment Discounted		
Period	sequence	value	(continued)	sequence	value	(continued)	sequence	value	(continued)	sequence	value	(continued)	sequence	value
0	-13000	-13000	13	-5892	-5051	26	-7300	-5365						
1	-5892	-5823	14	-3182	-2696	27	-7300	-5302						
2	-5892	-5754	15	-3182	-2664	28	-7300	-5240						
3	-5892	-5687	16	-3182	-2633	29	-7300	-5178						
4	-5892	-5620	17	-3182	-2602	30	-7300	-5117						
5	-5892	-5553	18	-3182	-2571	31	-7300	-5057						
6	-5892	-5488	19	-3182	-2541	32	-7300	-4997						
7	-5892	-5423	20	-3182	-2511	33	-7300	-4938						
8	-5892	-5360	21	-3182	-2481	34	-7300	-4880						
9	-5892	-5297	22	-3182	-2452	35	-7300	-4823						
10	-5892	-5234	23	-3182	-2423	36	-7300	-4766						
11	-5892	-5173	24	-3182	-2395	37	-7300	-4710						
12	-5892	-5112	25	-7300	-5429	38	278150	177342						
Sum of all discounted values (net present value)													0	
Sum of all savings (all negative values)													-219500	

Annex 2. Determining the annual interest rate using the simple return on savings and the age of the group

Using data from the group, one can look up the group's simple returns on savings and its age within the current cycle. The corresponding cell gives the effective annualised interest rate (%) resulting from standard financial calculations.

Simple returns on savings	Age of group this cycle												
	52	48	44	40	36	32	28	24	20	16	12	8	4
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2.5%	4.9%	5.4%	5.8%	6.4%	7.2%	8.1%	9.2%	10.8%	12.9%	16.2%	21.7%	32.9%	66.9%
5.0%	10.0%	10.8%	11.8%	13.1%	14.6%	16.5%	19.0%	22.3%	27.1%	34.5%	47.3%	75.1%	174.5%
7.5%	15.1%	16.4%	18.0%	19.9%	22.2%	25.3%	29.2%	34.6%	42.5%	54.9%	77.3%	128.8%	345.1%
10.0%	20.2%	22.0%	24.2%	26.9%	30.2%	34.4%	40.0%	47.8%	59.2%	77.7%	112.1%	196.7%	612.3%
12.5%	25.5%	27.8%	30.6%	34.1%	38.4%	44.0%	51.4%	61.8%	77.4%	103.0%	152.6%	282.0%	1025.7%
15.0%	30.8%	33.7%	37.2%	41.5%	46.9%	53.9%	63.3%	76.7%	96.9%	131.1%	199.3%	388.4%	1657.9%
17.5%	36.1%	39.6%	43.8%	49.0%	55.6%	64.2%	75.8%	92.5%	118.1%	162.1%	253.0%	520.5%	2614.0%
20.0%	41.6%	45.7%	50.6%	56.8%	64.6%	74.8%	88.9%	109.2%	140.8%	196.4%	314.6%	683.3%	4045.1%
22.5%	47.1%	51.8%	57.5%	64.7%	73.8%	85.9%	102.5%	126.8%	165.2%	234.0%	384.9%	883.0%	6165.6%
25.0%	52.6%	58.0%	64.6%	72.8%	83.3%	97.4%	116.8%	145.5%	191.4%	275.2%	464.8%	1126.7%	9277.9%
27.5%	58.3%	64.3%	71.8%	81.1%	93.1%	109.2%	131.7%	165.1%	219.4%	320.3%	555.5%	1422.5%	13804.5%
30.0%	64.0%	70.7%	79.1%	89.6%	103.2%	121.4%	147.2%	185.8%	249.4%	369.6%	658.0%	1780.1%	20331.0%

(Table continues on the next page)

Simple returns on savings	Age of group this cycle												
	52	48	44	40	36	32	28	24	20	16	12	8	4
32.5%	69.7%	77.2%	86.5%	98.2%	113.5%	134.1%	163.3%	207.6%	281.3%	423.2%	773.6%	2210.2%	29662.8%
35.0%	75.5%	83.8%	94.0%	107.0%	124.0%	147.1%	180.0%	230.4%	315.3%	481.5%	903.4%	2725.3%	42899.6%
37.5%	81.4%	90.5%	101.7%	116.0%	134.8%	160.5%	197.4%	254.3%	351.4%	544.8%	1048.9%	3339.9%	61532.7%
40.0%	87.3%	97.2%	109.5%	125.2%	145.9%	174.4%	215.5%	279.4%	389.7%	613.3%	1211.6%	4070.0%	87570.0%
42.5%	93.3%	104.0%	117.4%	134.6%	157.3%	188.6%	234.2%	305.7%	430.4%	687.3%	1393.0%	4934.3%	123698.7%
45.0%	99.3%	110.9%	125.4%	144.1%	168.9%	203.2%	253.6%	333.1%	473.4%	767.1%	1594.8%	5953.6%	173491.6%
47.5%	105.4%	117.9%	133.6%	153.8%	180.7%	218.3%	273.6%	361.8%	518.9%	853.1%	1818.7%	7151.5%	241671.8%
50.0%	111.6%	125.0%	141.8%	163.6%	192.8%	233.7%	294.3%	391.6%	567.0%	945.6%	2066.7%	8554.7%	334446.8%
52.5%	117.8%	132.1%	150.2%	173.6%	205.2%	249.6%	315.8%	422.8%	617.7%	1044.8%	2340.8%	10193.0%	459930.3%
55.0%	124.0%	139.3%	158.6%	183.8%	217.9%	265.9%	337.9%	455.2%	671.1%	1151.3%	2643.0%	12099.8%	628671.7%
57.5%	130.3%	146.6%	167.2%	194.2%	230.7%	282.6%	360.7%	489.0%	727.3%	1265.2%	2975.7%	14312.7%	854317.1%
60.0%	136.6%	153.9%	175.9%	204.7%	243.9%	299.7%	384.2%	524.1%	786.4%	1387.1%	3341.1%	16873.2%	1154431.4%
62.5%	143.0%	161.3%	184.7%	215.4%	257.3%	317.2%	408.5%	560.5%	848.5%	1517.2%	3741.8%	19828.0%	1551516.8%
65.0%	149.5%	168.8%	193.6%	226.3%	270.9%	335.1%	433.5%	598.4%	913.7%	1655.9%	4180.5%	23228.4%	2074271.5%
67.5%	155.9%	176.4%	202.6%	237.3%	284.9%	353.5%	459.2%	637.6%	982.0%	1803.6%	4659.9%	27131.7%	2759127.0%
70.0%	162.5%	184.1%	211.8%	248.5%	299.0%	372.2%	485.6%	678.3%	1053.6%	1960.8%	5182.9%	31601.1%	3652144.9%
72.5%	169.0%	191.8%	221.0%	259.8%	313.5%	391.4%	512.8%	720.4%	1128.4%	2127.9%	5752.6%	36706.4%	4811312.6%
75.0%	175.7%	199.5%	230.3%	271.3%	328.1%	411.1%	540.8%	764.0%	1206.8%	2305.2%	6372.2%	42524.4%	6309347.1%
77.5%	182.3%	207.4%	239.7%	283.0%	343.1%	431.1%	569.5%	809.2%	1288.6%	2493.2%	7045.2%	49139.8%	8237085.5%
80.0%	189.0%	215.3%	249.3%	294.8%	358.2%	451.6%	598.9%	855.8%	1374.0%	2692.3%	7775.0%	56645.6%	10707567.3%

(Table continues on the next page)

Simple returns on savings	Age of group this cycle												
	52	48	44	40	36	32	28	24	20	16	12	8	4
82.5%	195.8%	223.2%	258.9%	306.8%	373.7%	472.5%	629.2%	904.1%	1463.2%	2903.1%	8565.4%	65143.5%	13860963.2%
85.0%	202.6%	231.3%	268.6%	318.9%	389.4%	493.8%	660.2%	953.8%	1556.1%	3125.8%	9420.2%	74745.1%	17870472.4%
87.5%	209.4%	239.4%	278.5%	331.2%	405.3%	515.5%	692.0%	1005.2%	1652.9%	3361.2%	10343.6%	85572.2%	22949383.5%
90.0%	216.3%	247.5%	288.4%	343.6%	421.5%	537.7%	724.6%	1058.2%	1753.7%	3609.5%	11339.8%	97758.0%	29359475.3%
92.5%	223.2%	255.8%	298.4%	356.2%	437.9%	560.3%	758.0%	1112.9%	1858.5%	3871.3%	12413.1%	111447.2%	37421052.6%
95.0%	230.1%	264.1%	308.5%	368.9%	454.6%	583.3%	792.2%	1169.2%	1967.5%	4147.1%	13568.3%	126797.7%	47524724.3%
97.5%	237.1%	272.4%	318.8%	381.8%	471.6%	606.8%	827.1%	1227.3%	2080.8%	4437.4%	14810.1%	143981.0%	60145533.2%
100.0%	244.1%	280.8%	329.1%	394.9%	488.7%	630.7%	862.9%	1287.0%	2198.4%	4742.8%	16143.5%	163182.9%	75859401.6%

Annex 3. Determining the monthly interest rate using the simple return on savings and the age of the group

Using data from the group, one can look up the group's simple returns on savings and its age within the current cycle. The corresponding cell then gives you the effective monthly interest rate (%) resulting from standard financial calculations.

Simple returns on savings	Age of group this cycle												
	52	48	44	40	36	32	28	24	20	16	12	8	4
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2.5%	0.4%	0.4%	0.4%	0.5%	0.5%	0.6%	0.7%	0.8%	0.9%	1.2%	1.5%	2.2%	4.0%
5.0%	0.7%	0.8%	0.9%	0.9%	1.1%	1.2%	1.3%	1.6%	1.9%	2.3%	3.0%	4.4%	8.1%
7.5%	1.1%	1.2%	1.3%	1.4%	1.6%	1.7%	2.0%	2.3%	2.8%	3.4%	4.5%	6.6%	12.2%
10.0%	1.4%	1.5%	1.7%	1.8%	2.1%	2.3%	2.6%	3.1%	3.6%	4.5%	6.0%	8.7%	16.3%
12.5%	1.8%	1.9%	2.1%	2.3%	2.5%	2.8%	3.2%	3.8%	4.5%	5.6%	7.4%	10.9%	20.5%
15.0%	2.1%	2.3%	2.5%	2.7%	3.0%	3.4%	3.8%	4.5%	5.4%	6.7%	8.8%	13.0%	24.7%
17.5%	2.4%	2.6%	2.8%	3.1%	3.5%	3.9%	4.4%	5.2%	6.2%	7.7%	10.2%	15.1%	28.9%
20.0%	2.7%	2.9%	3.2%	3.5%	3.9%	4.4%	5.0%	5.8%	7.0%	8.7%	11.6%	17.2%	33.2%
22.5%	3.0%	3.3%	3.6%	3.9%	4.3%	4.9%	5.6%	6.5%	7.8%	9.7%	12.9%	19.2%	37.5%
25.0%	3.3%	3.6%	3.9%	4.3%	4.8%	5.4%	6.1%	7.2%	8.6%	10.7%	14.2%	21.3%	41.8%
27.5%	3.6%	3.9%	4.2%	4.7%	5.2%	5.8%	6.7%	7.8%	9.3%	11.7%	15.6%	23.3%	46.2%
30.0%	3.9%	4.2%	4.6%	5.0%	5.6%	6.3%	7.2%	8.4%	10.1%	12.6%	16.9%	25.3%	50.6%
32.5%	4.2%	4.5%	4.9%	5.4%	6.0%	6.8%	7.7%	9.0%	10.8%	13.6%	18.1%	27.3%	55.0%
35.0%	4.4%	4.8%	5.2%	5.8%	6.4%	7.2%	8.2%	9.6%	11.6%	14.5%	19.4%	29.3%	59.4%
37.5%	4.7%	5.1%	5.5%	6.1%	6.8%	7.6%	8.7%	10.2%	12.3%	15.4%	20.7%	31.3%	63.9%
40.0%	4.9%	5.4%	5.9%	6.4%	7.2%	8.1%	9.2%	10.8%	13.0%	16.3%	21.9%	33.2%	68.4%

(Table continues on the next page)

Simple returns on savings	Age of group this cycle												
	52	48	44	40	36	32	28	24	20	16	12	8	4
42.5%	5.2%	5.6%	6.2%	6.8%	7.5%	8.5%	9.7%	11.4%	13.7%	17.2%	23.1%	35.2%	72.9%
45.0%	5.4%	5.9%	6.5%	7.1%	7.9%	8.9%	10.2%	11.9%	14.4%	18.1%	24.3%	37.1%	77.5%
47.5%	5.7%	6.2%	6.7%	7.4%	8.3%	9.3%	10.7%	12.5%	15.1%	18.9%	25.5%	39.0%	82.1%
50.0%	5.9%	6.4%	7.0%	7.7%	8.6%	9.7%	11.1%	13.0%	15.7%	19.8%	26.7%	40.9%	86.7%
52.5%	6.2%	6.7%	7.3%	8.1%	9.0%	10.1%	11.6%	13.6%	16.4%	20.6%	27.9%	42.8%	91.3%
55.0%	6.4%	6.9%	7.6%	8.4%	9.3%	10.5%	12.0%	14.1%	17.0%	21.5%	29.0%	44.7%	96.0%
57.5%	6.6%	7.2%	7.9%	8.7%	9.6%	10.9%	12.5%	14.6%	17.6%	22.3%	30.2%	46.6%	100.6%
60.0%	6.9%	7.4%	8.1%	8.9%	10.0%	11.2%	12.9%	15.1%	18.3%	23.1%	31.3%	48.4%	105.3%
62.5%	7.1%	7.7%	8.4%	9.2%	10.3%	11.6%	13.3%	15.6%	18.9%	23.9%	32.4%	50.3%	110.1%
65.0%	7.3%	7.9%	8.6%	9.5%	10.6%	12.0%	13.7%	16.1%	19.5%	24.7%	33.5%	52.1%	114.8%
67.5%	7.5%	8.1%	8.9%	9.8%	10.9%	12.3%	14.2%	16.6%	20.1%	25.4%	34.6%	53.9%	119.6%
70.0%	7.7%	8.4%	9.1%	10.1%	11.2%	12.7%	14.6%	17.1%	20.7%	26.2%	35.7%	55.7%	124.4%
72.5%	7.9%	8.6%	9.4%	10.4%	11.5%	13.0%	15.0%	17.6%	21.3%	27.0%	36.8%	57.5%	129.2%
75.0%	8.1%	8.8%	9.6%	10.6%	11.8%	13.4%	15.4%	18.0%	21.9%	27.7%	37.8%	59.3%	134.0%
77.5%	8.3%	9.0%	9.9%	10.9%	12.1%	13.7%	15.7%	18.5%	22.4%	28.5%	38.9%	61.1%	138.9%
80.0%	8.5%	9.2%	10.1%	11.1%	12.4%	14.0%	16.1%	19.0%	23.0%	29.2%	39.9%	62.9%	143.7%
82.5%	8.7%	9.4%	10.3%	11.4%	12.7%	14.4%	16.5%	19.4%	23.6%	29.9%	40.9%	64.6%	148.6%
85.0%	8.9%	9.7%	10.6%	11.6%	13.0%	14.7%	16.9%	19.9%	24.1%	30.6%	42.0%	66.4%	153.5%
87.5%	9.1%	9.9%	10.8%	11.9%	13.3%	15.0%	17.3%	20.3%	24.6%	31.3%	43.0%	68.1%	158.4%
90.0%	9.3%	10.1%	11.0%	12.1%	13.5%	15.3%	17.6%	20.7%	25.2%	32.0%	44.0%	69.8%	163.4%
92.5%	9.4%	10.3%	11.2%	12.4%	13.8%	15.6%	18.0%	21.2%	25.7%	32.7%	45.0%	71.6%	168.3%
95.0%	9.6%	10.5%	11.4%	12.6%	14.1%	15.9%	18.3%	21.6%	26.2%	33.4%	46.0%	73.3%	173.3%
97.5%	9.8%	10.6%	11.6%	12.9%	14.3%	16.2%	18.7%	22.0%	26.8%	34.1%	47.0%	75.0%	178.3%
100.0%	10.0%	10.8%	11.9%	13.1%	14.6%	16.5%	19.0%	22.4%	27.3%	34.8%	47.9%	76.7%	183.3%

References

- Allen, H. and Panetta, D. 2010. Savings Groups: What Are They? Washington DC: The SEEP Network.
- Ardener, S. and Burman, S. (eds.) (1995) Money-Go-Rounds, Oxford: Berg.
- Brown, R.L. and Zima, P. 2011. Mathematics of Finance, 2nd ed. New York: McGraw-Hill.
- Collins, D., Morduch, J., Rutherford, S. and Ruthven, O. 2009. Portfolios of the Poor: How the World's Poor Live on \$2 a Day. Princeton, NJ: Princeton University Press.
- De Mel, S., McKenzie, D. and Woodruff, C. 2008. Returns to capital in microenterprises: evidence from a field experiment. *The Quarterly Journal of Economics*, 123(4), pp. 1329.
- EU Directive 2008/48/EC 1998.
- Graeber, D. 2011. Debt : the first 5,000 years. New York: Melville House Publishing.
- Homer, S. and Sylla, R.E. 1996. A History of Interest Rates. New Brunswick: Rutgers Univ Press.
- Ksoll, C., Lilleør, H.B., Lønborg, J.H. and Rasmussen, O.D. 2013. The Impact of Village Savings and Loans Associations on Household Welfare: Evidence from a Cluster Randomized Trial. In *The Microeconomics of Household Behavior in Sub-Saharan Africa*. PhD Thesis by Jonas Helth Lønborg, SDU, pp. 17-92.
- Lusardi, A. and Mitchell, O.S. 2007. Baby Boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of Monetary Economics*, 54(1), pp. 205-224.
- MFTransparency 2011. Formulas and Approaches Used to Calculate True Pricing.
- Rhyne, E. and Rippey, P. 2011. Crossfire: Formal vs. informal sector savings. *Enterprise Development and Microfinance*, 22(2), pp. 87-90.

Shipton, P.M.D. 2010. *Credit between Cultures: farmers, financiers, and misunderstanding in Africa*. New Haven: Yale Univ Press.

The Economist Dec 2011. Small wonder. A new model of microfinance for the very poor is spreading. The Economist, December 10th.

Truth in Savings Act 1991 FDIC IMPROVEMENT ACT OF 1991. Subtitle F—Truth in Savings.

Chapter 4

Walking the talk: the need
for a trial registry for
development interventions

Walking the talk: the need for a trial registry for development interventions

Ole Dahl Rasmussen

University of Southern Denmark and DanChurchAid

Thomas Barnebeck Andersen

University of Southern Denmark

Nikolaj Malchow-Møller

University of Southern Denmark

Abstract: Recent advances in the use of randomised control trials to evaluate the effect of development interventions promise to enhance our knowledge of what works and why. A core argument supporting randomised studies is the claim that they have high internal validity. The authors argue that this claim is weak as long as a trial registry of development interventions is not in place. Without a trial registry, the possibilities for data mining, created by analyses of multiple outcomes and subgroups, undermine internal validity. Drawing on experience from evidence-based medicine and recent examples from microfinance, they argue that a trial registry would also enhance external validity and foster innovative research.

Keywords: Impact assessment, randomised control trials, trial registry.

Acknowledgements: We thank two anonymous referees for helpful comments and suggestions.

This chapter has been published as a journal article. The full reference is: Rasmussen, O. D., N. Malchow-Møller, and T. B. Andersen (2011): "Walking the talk: the need for a trial registry for development interventions," *Journal of Development Effectiveness*, 3, 502-519. DOI: 10.1080/19439342.2011.605160

Introduction

In a recent column in the New York Times, two-time Pulitzer Prize winning journalist Nicholas D. Kristof celebrated the growing use of randomised control trials (RCTs) in development economics, calling it the ‘hottest thing in the fight against poverty’ (Kristof 2011, p. A27). RCTs, we are told, give us a good idea of what works in development aid. The present paper will argue that we are not quite there yet, but with a little effort we could come closer.

It is indisputable that RCTs have become a popular tool for assessing the impact of development interventions in recent years. There are several reasons for this, including general dissatisfaction with using regression techniques on non-experimental data (Freedman 1991), theoretical advantages of randomisation (Angrist and Pischke 2009, Banerjee and Duflo 2009), and perhaps even a general admiration of the hard sciences’ ideals. However, the approach has also received critique for having low external validity, and thus an inability to generalise findings to other contexts (Rodrik 2008, Deaton 2009), and for distorting the research agenda in development (Barrett and Carter 2010).

In this paper, we do not part with either side in the debate: Randomisation is indeed a promising approach for obtaining causal inference in the social sciences, although it comes with important limitations. Instead, we simply argue that current practice within the aid effectiveness literature in particular and development economics more generally often falls short of the stated aim of achieving high internal validity. To illustrate, simply note that researchers using RCTs can in principle choose between many different outcome measures. When a particular RCT study fails to credibly commit *ex ante* to a small number of outcome variables, we therefore have no way of knowing the amount of specification search that was required in order to uncover the reported finding.

To remedy this, we propose the establishment of a trial registry for development interventions with non-medical outcomes, which, among other things, allows researchers to commit to a specific outcome measure *ex ante*. A trial registry is different from a results database in

that changes made in the design and focus of the trial after the initial registration are clearly traceable. The idea of a trial registry for development interventions is not new. Duflo et al. have previously suggested the establishment of a database of studies that should also ‘include the salient features of the ex-ante design (outcome variables to be examined, sub-groups to be considered, etc.)’ (Duflo *et al.* 2007, p. 3910). Moreover, Ron Bose of the International Initiative on Impact Evaluation has suggested that the Consolidated Standards of Reporting Trials (CONSORT) for controlled medical trials should be adopted with some extensions to development interventions (Bose 2010). The CONSORT checklist specifies the pieces of information that should be included when reporting on medical RCTs. Item number 23 in the original CONSORT list is ‘Registration number and name of trial registry’ (Moher et al. 2010). Hence, adopting CONSORT would automatically call for a trial registry.⁴⁵

However, neither Duflo et al. (2007) nor Bose (2010) elaborate on these suggestions and, to the best of our knowledge, they have not been picked up by other researchers and/or practitioners in the field. We believe that the benefits of introducing a trial registry are both non-trivial and far too important to be mentioned only in passing. Or to put it differently: Without a trial registry, we fail to see how the randomisation approach, or other approaches claiming a high degree of internal validity, can credibly deliver on these promises of high internal validity.

We elaborate on these conjectures in the following. We also provide some specific suggestions for the most important features of such a registry and the institutional setup required for it to work. In doing so, we review literature and analyse data from evidence-based medicine and the largest trial registry within this field: ClinicalTrials.gov. Recent randomised studies from microfinance interventions are relied upon to illustrate our points. Finally, Appendix 1 provides a specific proposal for the contents of trial registry.

⁴⁵ Requiring authors reporting on RCTs to follow an augmented CONSORT would also lead to a much needed improvement in reporting on other issues, *e.g.*, the method used for the actual randomization (Bruhn and McKenzie 2009).

The immediate benefits of a trial registry

The implementation of a trial registry for development interventions has two immediate and desirable consequences. First, and most importantly, it works as a credibility enhancing mechanism. Second, it can increase external validity – particularly if the registry is extended to include results.

With respect to the first consequence, registering a trial in advance enhances credibility by improving internal validity. This may sound odd as the internal validity is usually considered the mainstay of RCTs. Hence, to illustrate, we (re-)consider the generally accepted advantage of RCTs compared to other types of studies: the increased ability to draw causal inference.

The issue at stake is attribution: how do we establish that the observed effect stems from our intervention and not from some excluded confounder? For non-experimental studies, a researcher will attempt to isolate the causal effect by controlling for a potentially long list of confounders or by using instrumental variable techniques. The obvious problem is that in theory it is impossible to know which confounders are relevant or which instruments are valid. To illustrate, a regression analysis of the connection between asbestos in drinking water and cancer controlled for a large number of background variables, but did not include smoking. Results were ‘highly statistically significant’, but for men only. Men are strong smokers and smoking thus appeared to be an unmeasured confounder (Kanarek et al. 1980, Freedman 1999).⁴⁶ At the same time, the researchers seemed to have carried out over 100 different specifications. This is what we in this context refer to as data mining: analysing a large number of subgroups and variables and reporting only the significant relationships.

⁴⁶ In this study smoking is a confounder only in the sense that smoking might have been correlated with asbestos levels by chance and we do not know this. It is difficult to imagine, as Freedman seems to suggest, that smoking causes higher levels of asbestos in drinking water as this comes from the level of natural occurring rock, serpentine, in the water reservoirs. Also, it is unlikely that areas with high asbestos in the drinking water would attract more smokers, as the level of asbestos is not known.

In a randomised study, on the other hand, a number of people or areas are randomly divided into two or more groups, and the intervention in question is implemented in only one of the groups. In this way, all confounders will in theory have the same distribution across the groups and the difference in the outcome measure stems from the intervention only. The researcher should not decide on a number of confounders to include or exclude, for which reason the possibilities for data mining are eliminated. Glaeser, for example, argues that using an experimental design limits the number of possible outcome variables which can be looked at, and thus makes data mining ‘essentially disappear’ (Glaeser 2006, p. 20). The attribution problem is solved, the data have spoken, and data mining is impossible.

Correct? Not entirely. Data mining can be a problem even in randomised studies. Most studies look at more than one outcome measure and often several subgroups. If conventional significance levels are used for the individual tests, there is a high chance of a type I error; that is, incorrectly rejecting a null hypothesis. In other words, studies of this kind are likely to find statistically significant effects on at least one of the outcomes, or for one of the subgroups, even in the absence of such effects, simply as a result of sampling variation. The point is simple to illustrate. Assume that we are testing m null hypotheses for m independent variables, representing either different subgroups or different outcome variables. Let α be the probability of a type I error for each of these tests; that is, the probability of falsely rejecting a true null hypothesis (finding a relationship where there is none). If the null hypotheses are all true, then the overall probability of a type I error (in other words, the probability of falsely rejecting at least one of the null hypotheses) is $\alpha_{overall} = 1 - (1 - \alpha)^m$. Hence, if $\alpha = 0.05$, then $\alpha_{overall} = 0.10$ if $m = 2$, while $\alpha_{overall} = 0.40$ if $m = 10$.⁴⁷ At the same time, studies are prone to various forms of

⁴⁷ With dependence between the variables, the picture becomes more complicated (and depends upon assumptions about the involved test statistics). Still, $\alpha_{overall} > \alpha$, unless the m variables are all perfectly correlated. As an example, assume that $m = 2$ and that the two test-statistics follow a bivariate standard normal distribution with a correlation coefficient of 0.5, then $\alpha_{overall} = 0.09$, and with a correlation coefficient of 0.9, $\alpha_{overall}$ still exceeds 0.07.

more or less conscious data mining. A researcher might thus report only on the subgroups and outcomes where there are significant results or just leave out certain insignificant findings. He or she might do so out of habit or because referees request more significant results as a prerequisite for publication.

These are purely mathematical operations; and whereas they do contribute to making data mining more difficult, it is still very much possible. The decision about which outcomes or subgroups to eventually include in a ‘family’ is left to the researcher. Hence it can be influenced more or less consciously by preliminary findings. A trial registry, where the relevant outcomes and subgroups are specified and registered ex-ante, would reduce the risk of data mining and allow others to at least gauge the extent to which the choice of outcomes and subgroups of interest have subsequently been changed, thereby improving internal validity of the findings. Basically, the solutions suggested previously places the internal validity of the results in the hands of the researcher.⁴⁸

The second immediate benefit of a trial registry is that it enables registration of all studies, regardless of whether results are positive, negative, statistically significant or insignificant. Hence, it makes available evidence less biased since information about non-published studies will allow for a more complete picture of the findings within a specific field, even if the trial registry does not contain the final results of the trial. Moreover, if a registry is extended to include results as well as the initial research design, the results from RCTs and other studies that find little or no effects are also accumulated. This can mitigate a publication bias where only significantly positive or negative effects are reported in journal articles, as argued by, for example, Glewwe and Kremer (2006). This is especially important in the case of RCTs, which have been criticised by several authors for their limited external validity; that is, the possibility of extrapolating the results to other settings and/or scaling the results to relevant

⁴⁸ To some extent, this is of course always the case. Any randomised experiment in social science must at some point rely on the subjective judgment of “experts” with local and/or context specific information (Cartwright 2007).

intervention levels (Deaton 2009). Yet the argument applies to all studies that rely on significance testing in one way or the other. An obvious way of improving the external validity is to accumulate as much evidence as possible from diverse settings and types of interventions. Only then can we hope to learn whether the findings from one study, RCT or not, are robust to changes in the settings and/or the level of intervention.

Is a trial registry really needed?

Is it possible that we do not need a trial registry for development interventions? One could argue that the degrees of freedom in the selection of outcome variables and subgroups are limited or that sufficient alternative measures have already been taken.

The flexibility in the choice of outcome variables and subgroups seems to be too large to be neglected. As an example, Karlan and Zinman note that there is no ‘natural summary statistic’ for household utility and thus choose to ‘measure treatment effects on a range of household survey variables that capture economic behaviour and subjective well-being’ (Karlan and Zinman 2010, p. 436). Also, the advance of initiatives like Measuring the Progress of Societies supported by OECD, European Union and United Nations shows that the choice of a single welfare indicator is not straightforward (OECD 2010). Finally, many RCTs in development use surveys with hundreds of questions to track outcomes. Consumption cannot be measured like blood pressure using a standard device, but requires questionnaire modules that are locally adapted and of considerable length, as the ones used in the Living Standard Measurement Surveys (Grosh and Glewwe 2000). Apart from questions about consumption, surveys commonly contain questions on assets, business activities, health, education and other potential outcomes. The final outcome measure is likely to reflect only a subset of the questions in the survey.

But could other steps do the trick? Glaeser argues in favour of making data publicly available and some journals have taken steps in this direction. One example is the American Economic Journal: Applied Economics, which has made it mandatory for submissions to

make data available together with calculation procedures to ensure the possibility of replicating the results. While this is certainly a good idea, it is no guardian against data mining related to the choice of outcome measures and subgroups along the lines described in the previous section. Moreover, and somewhat ironically, availability of data and computation procedures is likely to be more relevant in non-RCTs. Analyses using data from RCTs usually employ fairly simple regression frameworks, for which reason the exact programming of the data tends to matter less compared with non-experimental studies.

As described, data mining is possible in RCTs as well as in non-RCTs. Thus, a trial registry should be open to any survey-based research on development, regardless of whether it is based on an intervention that was randomly assigned or not.⁴⁹ The ex-ante registration of a trial or a survey ensures an explicit and visible distinction between primary and secondary analysis of data. Indeed, the largest trial registry for medical trials, ClinicalTrials.gov, includes all types of trials, whether randomised or not.⁵⁰

Secondary analysis is still important and should continue. But the possibility of increasing the credibility of the first hypothesis a dataset is used to test should not be foregone. When we mention RCTs specifically in this paper, it is because data for randomised trials is often collected with a specific research question in mind. Furthermore, RCTs frequently claim extraordinarily high internal validity, and the criticism raised in the present paper is therefore particularly relevant here.

Letting researchers research: other benefits of a trial registry

If a trial registry is implemented, unsubstantiated results will be less common and easier to detect. However, just as importantly,

⁴⁹ In principle, the registry could be open to any question involving statistical analysis, whether or not the question involves a survey. But since there is no way of assuring whether or not a researcher started the analysis prior to registration, this type of registration would not be credible.

⁵⁰ ClinicalTrials.gov terms this interventional and observational trials.

substantiated yet surprising and small effects would also be easier to spot. Consider the situation without a trial registry: when reviewers of academic journals receive a paper reporting on a randomised trial, they need to judge whether the reported results are genuine or whether they could be a result of data mining. Most probably, they will look at whether the outcome measures in question are obvious choices as dependent variables. The basis for doing this is the reviewers' sense of the field and the received knowledge in the field; possibly together with their own understanding of the causality in question. Conversely, with a trial registry in place, this decision can be made by the researcher in advance, who is therefore free to choose the outcome variable(s) of interest, allowing her to include more novel outcomes and hence more uncommon variables according to the received wisdom in the field. The registry then provides the a priori credibility of these outcome variables, not the common sense of the reviewers.

Recent evidence in microfinance provides an illustrative example: Banerjee et al. (2010) carry out a randomised experiment on access to microfinance; they find no effect on female empowerment and total consumption, which in the general debate are typically imagined outcomes of microfinance. At the same time, Banerjee et al. find a negative effect on spending on temptation goods and a positive effect on spending on durables. Also, the number of people opening businesses was 1.7 percentage points higher in the treatment areas compared with the control areas, an increase from 5.3 per cent to 7.0 per cent. One in five of the loans that came from the opening of new microfinance branches resulted in starting a new business (Banerjee et al. 2010). The problem with these results is not so much that the effects are small, but that the outcome variables are non-trivial. In the light of standard economic theory, the expectation would be that increased credit availability leads to business start-ups for the previously credit constrained clients. In this view, a rise of 1.7 percentage points seems to be a very small effect and the absence of effects on consumption and female empowerment is worrying.

But there are other frameworks for understanding microfinance. Microfinance can be seen as a commitment device for sophisticated

agents with time-inconsistent preferences (Ashraf et al. 2006). These agents have a tendency to spend income immediately when they receive it, but because they are aware of this, it is possible for them to commit to not spending future income. Microfinance loans and savings products serve as commitment devices. Within this framework, finding that spending is moved from temptation goods to investment goods is an interesting result and a rise in business start-ups of 1.7 percentage points might be quite a change.

Thus, the two frameworks yield two different interpretations and predictions. If the relevant outcome measures are decided upon by a broad range of social science researchers, the advance of new and surprising findings is unlikely. In our opinion, this will be the case when there is no trial registry in place. With a trial registry, researchers can decide upon the framework themselves and new and ground-breaking evidence will stand a better chance of getting acknowledged.

Credibility of non-standard outcomes is likely to be important in assessing the impact of development interventions, in particular those with non-medical outcomes, due to the choice of treatment units and the time perspective involved. These two factors constrain the power of the studies, a fact that is widely acknowledged in the literature on community-based epidemiology (Sorensen et al. 1998, Atienza and King 2002). When the treatment unit is communities, schools or districts, the take-up rate is important for the power of the study. The above study by Banerjee et al. provides a case in point: opening microfinance branches in randomly selected areas led to an increase in borrowing from microfinance institutions of a mere 8.3 percentage points (Banerjee et al. 2010) compared to the control areas. Any difference in outcomes must be driven by this difference in take-up rates. In other words, the intervention analysed is an ‘intention to treat’ the individuals in the selected areas. With low compliance in the treated areas, the effect on those few who actually comply with the treatment (in other words, borrow from the institution) must be correspondingly larger for the study to show an effect, compared with a typical medical situation, where individual randomisation usually

leads to compliance rates over 75 per cent. In an instrumental variable framework, this is similar to a low first stage; that is, a weak instrument, and potential issues of small sample bias are well known (Murray 2006).

The long-time perspective often involved is another constraining factor. A microfinance intervention might relax credit constraints and increase activity, but only after clients have learned about and understood the products, found a business idea or expanded their business and started selling. This is likely to take time. If the effect is instead an increased ability to smooth consumption, then the effect on consumption levels is likely to take even longer. One solution is to wait for mechanisms to work, something which might take two, five or 10 years. Another is to track immediate effects, for example business investments or consumption smoothing and then simulate the effects on consumption or empowerment (for example Sorensen et al. 1998). This, however, requires credibility in non-standard outcome measures, something that is difficult in the absence of a trial registry.

Copying the wheel: trial registries in medicine

To the best of our knowledge, medicine is the only field to use trial registries at present. Several other fields have databases that are sometimes called trial registries, but since they do not allow for identification of changes made in the design or focus of the trial during the implementation and the subsequent analysis of the data, we do not consider them here.⁵¹ Hence, instead of re-inventing a trial registry for development interventions from scratch, it may be worthwhile to consider some of the arguments and evidence advanced in favour of a trial registry in medicine. Furthermore it may be instructive to draw on the experience from medicine in implementing such a registry.

A key reason behind trial registries in medicine is the belief that registering trials will reduce the bias towards statistically significant

and positive findings in the available evidence (Ioannidis 2005, Zarin et al. 2007). A potential consequence of such a bias is that wrong or harmful treatments of diseases are continued despite the existence of unpublished relevant evidence (Dickersin and Rennie 2003). This is unethical both towards patients and towards participants in trials who often volunteer because they believe that they contribute to the advancement of medical knowledge.

A trial register can reduce the possibility of researchers changing their hypotheses during the course of a study. One reason for changing outcomes could be that more statistically significant studies get published more easily. Easterbrook et al. (1991) report that the probability of publishing increases if studies have statistically significant results. The time from trial registration to publication has also been shown to be negatively affected by the level of statistical significance of the results (Ioannidis 1998). Also, funders of studies may put more effort into publishing certain results over others (Dickersin 1990). Such studies could, however, themselves be subject to selection bias: confounding factors like researcher age or ability might cause both statistical power and higher publication probability, for example due to research design.

If we briefly turn our attention outside medicine, we find an interesting experiment in cognitive psychology that points toward the existence of a publication bias. Mahoney (1977) implemented a randomised trial changing only the results section in a paper to 78 referees. The paper studied whether or not extrinsic rewards leads to changes in behaviour, a controversial topic at the time. The referees' evaluations of the paper's publication merit, but also of the quality of the methods section, increased with the direction and strength of the results. That not only the significance but also the direction of the results may matter is supported by Simes (1986) who found significant differences between published and unpublished RCTs registered in a trial registry for cancer research: published studies were

⁵¹ These registries include The Cochrane Central Register of Controlled Trials (CENTRAL) and the The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR).

more positive. Rasmussen et al. (2009), however, did not find any differences in results for studies on a specific cancer drug.

A trial registry is not necessarily an effective tool against publication bias. That will depend on the reasons underlying the publication bias. A trial registry will only affect publication bias in cases where the bias stems from the fact that authors are changing outcomes of a study. Changing outcomes, under-reporting outcomes or simply not reporting outcomes seems to have been common in medicine, at least during the 1990s. Chan et al. (2004b) found that 40 per cent of published studies funded by the Canadian Institutes of Health Research had changed their primary outcome, whereas Chan et al. (2004a) found at least one unreported outcome in 50–65 per cent of published studies based on all medical trials approved by the Scientific-Ethical Committees for Copenhagen and Frederiksberg, Denmark, between 1994 and 1995. Eighty-six per cent of authors responding to a questionnaire denied in the same study to have unreported outcomes. Finally, authors might not finalise studies that appear to show no statistical significance (Ioannidis 1998). Registering trials in one primary registry became widespread only after 2005, as the next section describes.

Apart from providing a disincentive to change outcomes, a trial registry will also make it easier to assess the magnitude of bias in the available evidence. Without a registry, data on the comparison group that consists of planned and unpublished studies, research designs and results are difficult data to compile.⁵² Because of this, studies looking at publication bias have used trial registries (Milette et al. 2011) or protocols submitted to Ethical Committees (Chan et al. 2004a, Chan et al. 2004b) as a basis for comparing published and non-published results.

Several of the issues raised in medicine are relevant for development interventions. With respect to the registry's function as a credibility enhancing mechanism, the arguments and evidence from

⁵² When it comes to results, a trial registry will not necessarily make data collection easier as many existing trial registries do not contain results, for example ClinicalTrials.gov which is the largest existing registry.

medicine would most probably apply in development: a functioning trial registry will discourage researchers from changing outcomes during a trial, and it will encourage authors to report the initially chosen outcomes. This is only reinforced if the registry is extended to include results: a trial registry will make it easier to assess the magnitude of a potential publication bias by making evaluations with negative or insignificant findings available to practitioners and the research community.

A recent review of evidence on microcredit found that all except one of the evaluations carried out by donor agencies and large non-governmental organisations showed positive and significant effects, suggesting that bias exists (Kovsted et al. 2009). Likewise, publication bias within academic fields that publish on development interventions is likely to exist. In the absence of a trial registry, authors have analysed publication bias indirectly by looking at the distribution of significance levels of published results. Many have found these levels to be more positive than what is realistic. As such, DeLong and Lang (1992) show that less than one third of studies reporting a significantly positive result in economics are likely to be true. Similar results are found in political science (Gerber et al. 2010).

A relevant question is whether trial registration is a cost effective way of enhancing credibility; after all, credibility can be achieved in many other ways, for example by adjusting p-values, increasing power through more and better data or by increasingly replicating already published studies. Given the cost of data collection and the nature of the service provided by a trial registry as a public good, it is likely to be cost effective, but an actual analysis of this issue would be justified.

In medicine, trial registries with results are used for writing systematic reviews and conducting meta-studies (Higgins and Green 2006). Systematic reviews have recently started to appear in development together with more traditional literature reviews, for example looking at the effects of microfinance (Kovsted et al. 2009, Stewart et al. 2010). For this exercise to be meaningful in development, however, the studies reviewed must be relevant for

other populations than the one being studied; that is, they must have external validity. In medicine this is less of a problem since many treatments can be described and replicated, and people can be expected to react in similar ways to a drug whether they live in the slums of Hyderabad or the countryside of California. For development, the case is different. Treatments are not easily described since they usually involve institutional setups that are specific to the country and context in question. Also, people cannot be expected to react in similar ways to microfinance in two different countries, economies and cultures.⁵³ To use the language often invoked to justify randomised trials, unit homogeneity might not be necessary to recover the average treatment effect (Holland 1986), but it is necessary for external validity. For these reasons, the usefulness and content of a trial registry with results on development is likely to differ from one in medicine. In particular, a trial registry for development should contain descriptive information about the treatment and the population that is rich enough to give an impression of how and in which context the implementation was carried out. In other social science fields, for example in the study of welfare programmes, it is common to supplement the evaluation with so-called implementation reports in an attempt to accomplish this (for example Kingwell et al. 2005).

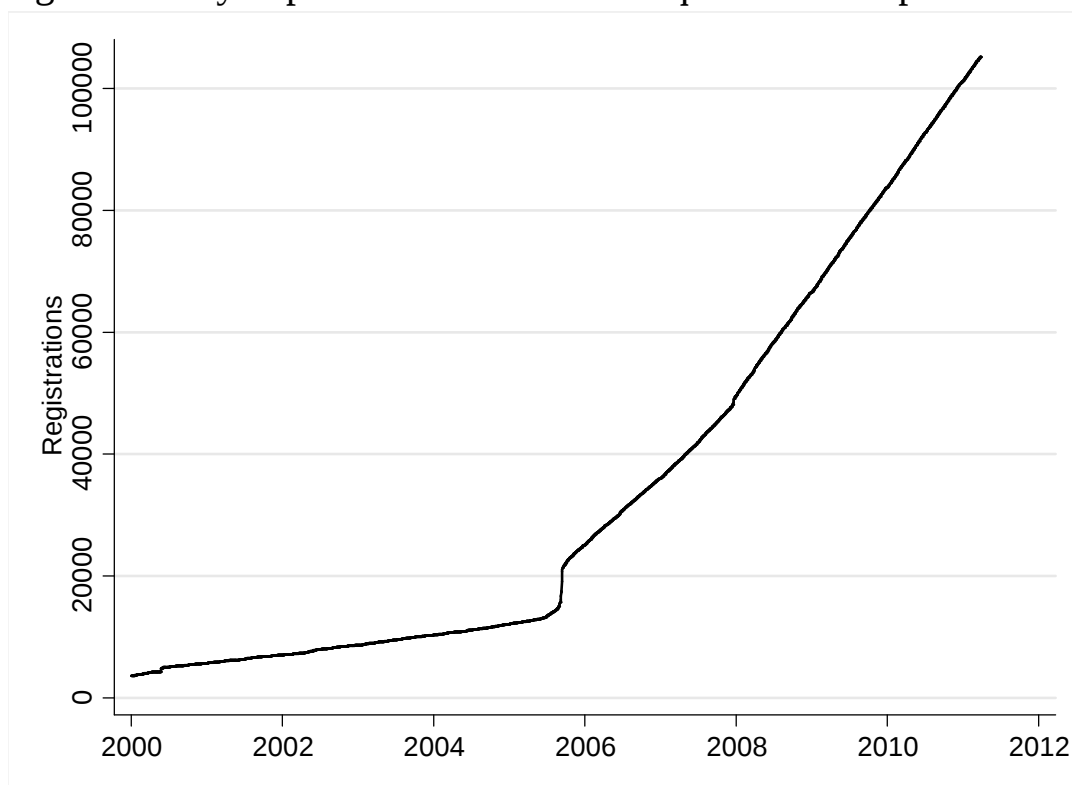
Experience from medicine suggests that getting researchers to register trials can happen quickly, but it requires that the necessary structures are in place; and that may take time. In 2006 the Cochrane Collaboration reported that ‘no single, central registry of ongoing randomised trials currently exists.’ Since then things have changed: ClinicalTrials.gov, which is now the leading trial registry for studies related to health, has over 100,000 registered records, up from 7000 on 1 January 2002 (cf. Figure 1). This seems to be driven by two primary factors: passing of legislation from 1997 onward that requires registration prior to approval of drugs and the requirement from

⁵³ The study by Coleman (1999) is a case in point. Using a pipeline method approach for identification, Coleman finds no effect of microcredit in Northern Thailand. He then notes that this negative finding may be due to an abundance of credit in the region, for which reason the results may have little external validity.

journal editors that studies will only be accepted for publication if they have registered their trial prior to admitting the first person to treatment (Zarin et al. 2005, Zarin et al. 2007). The latter is required by the more than 850 journals following the Uniform Requirements for Manuscripts issued by the International Committee of Medical Journal Editors (ICMJE) (De Angelis et al. 2004). Indeed, statistical analysis as well as visual inspection of the number of registrations at ClinicalTrials.gov over time suggests that the adoption of these requirements on 13 September, 2005, indeed had an effect (Figure 1) (Zarin et al. 2005).

Figure 4. Total number of trial registrations at ClinicalTrials.gov.

International Committee of Medical Journal Editors established registration by September 13 2005 as a requirement for publication.



Upon inspection of Figure 1, it would appear that the policy of ICMJE had a large effect. The average number of registrations per day before this date is different from the average number of registrations after this date at a statistically significant level (the difference is 30.0 with a

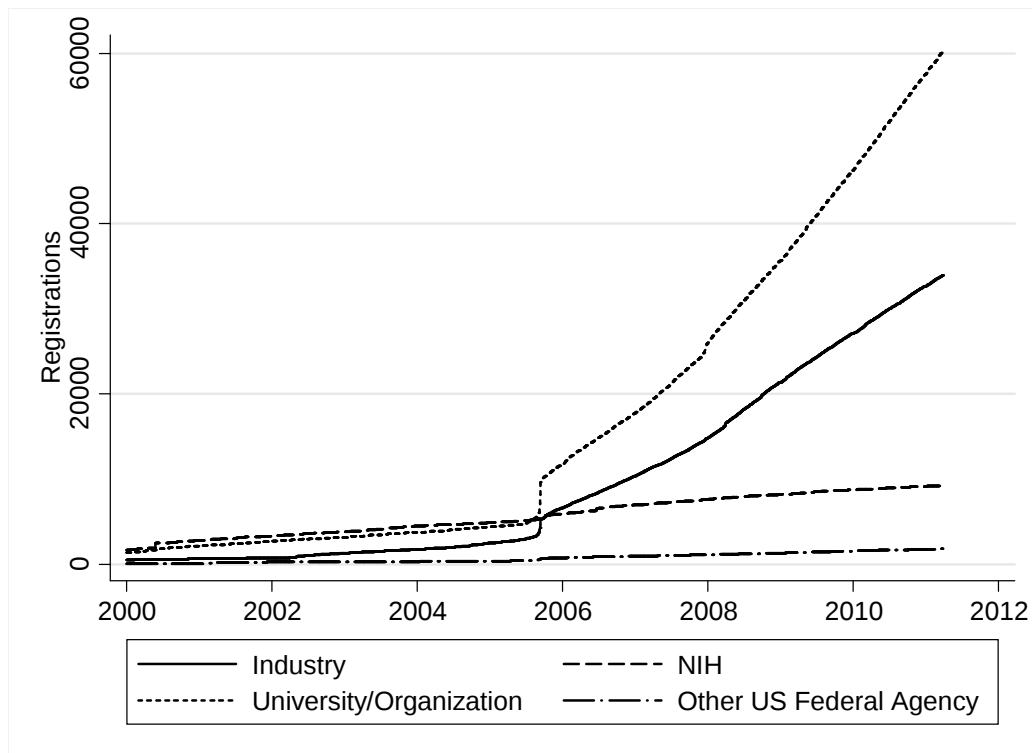
confidence interval between 26.50 and 33.54). The threat of rejection from journals is effective. Making use of this fact in the field of development interventions, however, requires that we look beyond the averages. Not everyone responded to the policy in the same way. Indeed, trials funded by different donors responded differently: Studies funded by the industry or universities were responsible for the largest share of the change in the rate of registration. In contrast, the rate of registration did not change for studies funded by the National Institute of Health (NIH) or other federal agencies (cf. Figure 2). The average number of daily of registered trials funded by NIH actually decreased by a small but statistically significant amount, whereas the change in daily registrations for trials funded by other federal agencies did not change at a 5 per cent level (see table 1). The fall in registrations is not immediately visible in Figure 2, since the figure displays cumulative data.

Table 1: Average daily registrations - differences in means

	Number of days	Before Sep 13 2005	After Sep 13 2005	Difference	95% Confidence interval for the difference	
Industry	3293	3,52	14,60	11,08	10,01	12,16
Universities	3293	6,03	25,84	19,82	17,51	22,12
NIH	3293	2,92	1,89	-1,03	-1,79	-0,28
Other federal agency	3293	0,45	0,60	0,15	-8,33E-04	0,30

Two explanations seem possible: NIH-funded authors might not be interested in publication or NIH-funded authors registered their trials already before September 2005. Within the field of development, development agencies might not be interested in publication and thus a policy by journal authors might not affect them.

Figure 5. Looking beyond the averages: The policy by International Committee of Medical Journal Editors did not affect all.



Experience from medicine extends beyond the trial registry itself. To facilitate the dissemination of trial results, funders, editors and publishers who promote RCTs could agree on a set of minimum standards for reporting similar to the CONSORT Checklist (Moher et al. 2010), which has been documented to increase quality in reporting in medicine (Plint et al. 2006). An adaption has thus been suggested within political science (Boutron et al. 2010). The work by Bose (2010), could provide a starting point for this work.

Creating a credibility infrastructure: a trial registry is not enough

Being merely a database of records, a stand-alone trial registry is unlikely to reduce the probability of type I errors by itself. If the trial registry is not supported by systems that give researchers clear incentives to register, then it is unlikely to have any effect. The fact that registries have existed in medicine for more than 50 years, but have only been widely used during the past decade, illustrates the

importance of getting structures right. The examples from medicine provided above could be mimicked in development.

Funders like bilateral and multilateral donors as well as foundations could make trial registration compulsory in contracts with evaluators. Supervisors of funders, like the OECD, could facilitate common agreement among funders on this. The OECD already enjoys considerable legitimacy in the realm of evaluation in that many development agencies follow the organisation's guidelines. Moreover, key journals could make registration a prerequisite for publishing. To avoid collective action problems, momentum would have to be created to make several important journals commit at the same time. In the process organisational and technical experiences from medicine should be taken into account (see for example McCray and Ide 2000). Furthermore, for the registry to have an effect, someone must do the actual comparison of the original hypotheses with the actually tested hypotheses. This endeavour is likely to be undertaken by researchers in developing country governments, development agencies or universities.

Finally, a trial registry has several limitations. Even a well-functioning registry will not remove type 1 errors. On the contrary, there is a risk that a registry installs a false sense of security in studies' results. Data mining is still possible as published outcomes can be different from registered ones, publication bias can still happen if journals accept studies on the basis of statistical significance, and there might still be a bias in the non-published available evidence since results of negative or statistically insignificant studies may still not be made public. A trial registry is unlikely to solve all the problems we have concerning bias. But it is likely to reduce some of them, and for that purpose we believe that it is worth doing.

What are the important features of a trial registry?

To effectively perform the functions described above, a trial registry needs to have certain features. Below we describe what we think are the most important ones, many of which are adapted from the

requirements set by WHO for trial registries in the medical field (WHO 2010).

Tracking changes. By far the biggest difference between a trial registry and a database of results is the ability to track changes in the choice of outcome variables since the initial registration. ClinicalTrials.gov contains information on the last update as well as side-by-side display of changes made.

Content. The registry should be open to all prospective registrants, should publically display key information of the trial and should never delete a registered trial. It should at least contain the following information about each trial: a unique ID number, registration date, sources of monetary support, contact details, country, intervention, type of study, date of first enrolment, primary outcome(s), secondary outcomes and planned subgroup analyses.⁵⁴ Appendix 1 for a more elaborate suggestion.

Unambiguous identification. The registry should have a process for identifying double registrations and it should be linked to other registries within other fields to enable cross checks. This is important to avoid multiple ex-ante entries of the same trial where only the most suitable entry is evoked ex-post.

Governance. To ensure that the trial registry is perceived as legitimate, it should be governed by a board with broad representation from actors with a stake in impact assessments. Representation should be across low-income and high-income countries and should include researchers, evaluators and policy-makers.

The development sector already has some databases of randomised trials (e.g. International Initiative for Impact Evaluations n.d., OECD's DERE n.d.), but these do not meet the list of requirements above and

⁵⁴ This list was inspired by the WHO Trial Registration Data Set version 1.2.1.

do not allow for clear identification of the original registration by making changes clearly visible. As mentioned above, the same is true for registries created within the fields of education (What Works Clearinghouse 2010) and criminology (C2-SPECTR 2010). Needless to say, the institutions cited who currently maintain databases of randomised trials for development, like the OECD and the International Initiative for Impact Evaluations, would be good candidates as hosts for a trial registry.

Conclusions

There is a growing recognition that recent advances in the use of randomisation as a means to assessing the effect of development interventions has contributed importantly to our knowledge of what works and why. By emphasising causality, it has enabled researchers to provide evidence of effects as well as advances in theory development; for example, within development economics. In doing this, proponents have pledged allegiance to hard scientific ideals and claimed a ‘credibility revolution in development economics’ (Angrist and Pischke 2010, p. 3).

We argue that development work and development economics still lack much of the infrastructure that is needed to generate this type of credibility. If randomised studies of development interventions are to stay true to their claims of high internal validity, we need to establish a bulwark against the data mining options that pose a threat to credibility also for randomised studies. A trial registry will enable readers and reviewers to judge whether the results reported were decided prior to the study, whether from a randomised control trial or another type of study, and thus help in interpreting the extent to which reported results could be a consequence of changes made to outcome measures in question. Results on outcome variables and subgroups not mentioned in the ex-ante registration of the study should still be reported, but with a registry it will be possible to distinguish primary outcomes from secondary outcomes. Apart from making the claim of internal validity more trustworthy, a trial registry would also increase external validity by facilitating comparisons of trials across different

contexts. Finally, a trial registry would place the decision of relevant outcome variables solely with the researcher, allowing her to include novel outcome measures, while still being able to report results credibly. In this way, a trial registry would promote innovation in theory and practice.

Agreeing that a trial registry is a good idea is likely to be the easiest part of actually implementing one. Participation from relevant stakeholders, researchers and practitioners is arguably much more difficult. Experience from medicine shows that it is important to get incentives right for researchers to start registering and we suggest that donors and journal editors should set a date after which registration is required for grants and papers. A prerequisite for mobilising support for the idea is a clear understanding of its benefits for research, evaluation, aid and policy making. Moreover, learning from the experience within medicine can increase the chances of success. In this paper, we have put forth what we think are the primary reasons why a trial registry is indeed needed for development interventions with non-medical outcomes. By doing this we hope to have moved the effective implementation of a trial registry one step closer.

Appendix 1. Proposal for contents of a trial registry for development interventions

For a trial registry to be effective, major stakeholders must decide among themselves the required contents and the format of the registry. Below is a list of items that should be considered in that regard; the list draws inspiration from the WHO Trial Registration Data Set version 1.2.1.

In the below suggestion, a trial registry for development interventions contains three types of information about trials: general trial information, information on changes, and study results. General trial information includes compulsory information which is available before results of the trial are available. Information on changes is an automatically generated list of changes which have been made to the general trial information. Study results remain optional.

Pre-result information

Unique ID number

Registration date

Sources of monetary support

Major sources of monetary or material support for the study, for example university, foundation, government or company.

Primary Sponsor

This is the individual, organization, group, or other legal entity which takes responsibility for the initiation and management of the study. The Primary Sponsor is responsible for ensuring that the trial is properly registered. The Primary Sponsor may or may not be the main funder.

Local Sponsor

If the Primary Sponsor is registered in a different country than where the trial takes place, a local sponsor should be assigned. This can be a local research body involved or an implementing agency.

Location of the study

Which country does the study takes place.

Intervention

For each arm of the trial, provide a brief description of the intervention.

Type of study

Interventional, random assignment

Non-interventional, quasi-experimental

Non-interventional, other

Other typologies include the ones suggested by Charles S. Reichart (2011), by John List on fieldexperiments.com or 3ie.org's results database.

Date of first enrolment

Date when the first person started or will, according to plan, start participation in the intervention.

Target Sample Size

- Number of randomized units that the study plans to study, e.g. persons, villages, schools.
- Number of participants involved in the study
- Number of studied participants: The number of people who are interviewed or otherwise provided data on.

In a simple design, the three sample size figures can be the same.

Primary outcome(s)

Outcomes are events, variables, or experiences that are measured because it is believed that they may be influenced by the intervention.

The Primary Outcome should be the outcome used in sample size calculations, or the main outcome(s) used to determine the effects of the intervention(s). Most trials should have only one primary outcome.

For each primary outcome include:

- The name of the outcome
- The metric or method of measurement used

- The time point(s) of primary interest

Secondary outcomes

Secondary outcomes are outcomes which are either of secondary interest or that are measured at time points of secondary interest. A secondary outcome may involve the same event, variable, or experience as the primary outcome, but measured at time points other than those of primary interest.

As for primary outcomes, for each secondary outcome provide:

- The name of the outcome
- The metric or method of measurement used
- The time point(s) of interest

Subgroup analyses

Which subgroups does the study plan to analyse.

History of changes

This format is inspired by ClinicalTrials.gov.

Changes to ID 348273

	Before 5/7/2010	After 2/1/2011
Primary outcomes	Total consumption as measured by a survey of all food and non-food items	Total consumption as measured by a survey of all food

Study results

Participants flow

In the case of an interventional study with random assignment: How many participants or units were selected for the study, and how many completed?

For example

	Treatment Group	Control
STARTED	450	226
COMPLETED	418	205
NOT COMPLETED	32	21
Withdrawal by Subject	16	11
Lost to Follow-up	16	10

Baseline Characteristics

Key baseline characteristics

Outcome measures

Information on the originally selected primary and secondary outcomes.

Implementation details

Did the implementation go as planned? Where can interested organisations find additional information on how to implement a similar intervention?

Context of study

Which contextual factors can have affected the study, including political, cultural and economic contextual factors?

Publications

References, if the results were published.

References for this appendix

Reichardt, C.S., 2011. Evaluating Methods for Estimating Program Effects. *American Journal of Evaluation*, 32 (2), 246-272.

WHO Trial Registration Data Set version 1.2.1:

<http://www.who.int/ictrp/network/trds/en/index.html>.

References

- Angrist, J. and Pischke, J. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *The Journal of Economic Perspectives*, 24(2), pp. 3-30.
- Angrist, J.D. and Pischke, J.-S. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.
- Ashraf, N., Karlan, D. and Yin, W. 2006. Tying Odysseus to The Mast: Evidence From a Commitment Savings Product in the Philippines. *Quarterly Journal of Economics*, 121(2), pp. 635-672.
- Atienza, A.A. and King, A.C. 2002. Community-based health intervention trials: an overview of methodological issues. *Epidemiologic reviews*, 24(1), pp. 72-79.
- Banerjee, A. and Duflo, E. 2009. The Experimental Approach to Development Economics. *Annu. Rev. Econ*, 1(1), pp. 151-178.
- Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. 2010. *The miracle of microfinance? Evidence from a randomized evaluation*. Working Paper: Financial Access Initiative and Innovations for Poverty Action.
- Barrett, C.B. and Carter, M.R. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy*, 32(4), pp. 515.
- Bose, R. 2010. CONSORT Extensions for Development Effectiveness: guidelines for the reporting of randomised control trials of social and economic policy interventions in developing countries. *Journal of Development Effectiveness*, 2(1), pp. 173-186.
- Boutron, I., John, P. and Torgerson, D.J. 2010. Reporting Methodological Items in Randomized Experiments in Political Science. *The ANNALS of the American Academy of Political and Social Science*, 628(1), pp. 112-131.

- Bruhn, M. and McKenzie, D. 2009. In pursuit of balance: randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4), pp. 200-232.
- C2-SPECTR 2010. *The Campbell Collaboration Trials Register (C2-SPECTR)*: Available from: <http://geb9101.gse.upenn.edu/RIS/RISWEB.ISA> [Accessed 29 October 2010].
- Cartwright, N. 2007. Are RCTs the gold standard? *Biosocieties*, 2(1), pp. 11-20.
- Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C. and Altman, D.G. 2004a. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*, 291(20), pp. 2457-2465.
- Chan, A.W., Krleza-Jeric, K., Schmid, I. and Altman, D.G. 2004b. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal*, 171(7), pp. 735-740.
- Coleman, B.E. 1999. The impact of group lending in Northeast Thailand. *Journal of Development Economics*, 60(1), pp. 105-141.
- De Angelis, C., Drazen, J.M., Frizelle, F.A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A. and Overbeke, A. 2004. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Annals of internal medicine*, 141(6), pp. 477.
- Deaton, A. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Cambridge, MA.: NBER Working Paper 14690, National Bureau of Economic Research.
- DeLong, J.B. and Lang, K. 1992. Are all economic hypotheses false? *The Journal of Political Economy*, 100(6), pp. 1257-1272.

- Dickersin, K. 1990. The existence of publication bias and risk factors for its occurrence. *JAMA*, 263(10), pp. 1385.
- Dickersin, K. and Rennie, D. 2003. Registering clinical trials. *JAMA*, 290(4), pp. 516.
- Duflo, E., Glennerster, R. and Kremer, M. 2007. Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4, pp. 3895-3962.
- Easterbrook, P.J., Gopalan, R., Berlin, J. and Matthews, D.R. 1991. Publication bias in clinical research. *The Lancet*, 337(8746), pp. 867-872.
- Freedman, D. 1991. Statistical models and shoe leather. *Sociological methodology*, 21, pp. 291-313.
- Freedman, D. 1999. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14(3), pp. 243-258.
- Gerber, A.S., Malhotra, N., Dowling, C.M. and Doherty, D. 2010. Publication Bias in Two Political Behavior Literatures. *American Politics Research*, 38(4), pp. 591.
- Glaeser, E. 2006. *Researcher Incentives and Empirical Methods*. Cambridge, MA.: NBER Technical Working Paper No. 329, National Bureau of Economic Research.
- Glewwe, P. and Kremer, M. 2006. Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, 2, pp. 945-1017.
- Grosh, M. and Glewwe, P. 2000. Designing household survey questionnaires for developing countries : lessons from 15 years of the Living Standards Measurement Study. Washington DC, 3 volumes: World Bank.
- Higgins, J. and Green, S. 2006. Locating and selecting studies. In H. Jpt and G. S (eds.) *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]*. Chichester, UK: John Wiley & Sons, Ltd. .
- Holland, P.W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), pp. 945-960.

- International Initiative for Impact Evaluations n.d. *Database of Impact Evaluations*.
http://www.3ieimpact.org/database_of_impact_evaluations.html accessed October 2010: International Initiative for Impact Evaluations.
- Ioannidis, J. 1998. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*, 279(4), pp. 281.
- Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8), pp. e124.
- Kanarek, M., Conforti, P., Jackson, L., Cooper, R. and Murchio, J. 1980. Asbestos in drinking water and cancer incidence in the San Francisco Bay Area. *American Journal of Epidemiology*, 112(1), pp. 54-72.
- Karlan, D. and Zinman, J. 2010. Expanding credit access: Using randomized supply decisions to estimate the impacts. *Review of Financial Studies*, 23(1), pp. 433-464.
- Kingwell, P., Dowie, M., Holler, B. and Vincent, C. 2005. *Design and Implementation of a Program to Help the Poor Save*. Canada: SDRC.
- Kovsted, J., Andersen, T.B. and Kuchler, A. 2009. *Synthesis of impact evaluations of microcredit*. Copenhagen: Danish Ministry of Foreign Affairs.
- Kristof, N.D. 2011. Getting Smart on Aid. *New York Times*, May 18.
- Mahoney, M.J. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1(2), pp. 161-175.
- McCray, A.T. and Ide, N.C. 2000. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association*, 7(3), pp. 313-323.
- Milette, K., Roseman, M. and Thombs, B.D. 2011. Transparency of outcome reporting and trial registration of randomized controlled trials in top psychosomatic and behavioral health journals: A systematic review. *Journal of Psychosomatic Research*, 70(3), pp. 205-217.

- Moher, D., Hopewell, S., Schulz, K., Montori, V., Gotzsche, P., Devereaux, P., Elbourne, D., Egger, M. and Altman, D. 2010. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, 340, pp. c869.
- Murray, M.P. 2006. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4), pp. 111-132.
- OECD's DEREc n.d. *DAC Evaluation Ressource Centre*. Paris: OECD. Available from: http://www.oecd.org/pages/0,2966,en_35038640_35039563_1_1_1_1_1,00.html [Accessed 29 October 2010].
- OECD 2010. *Istanbul Declaration*. Istanbul, <http://www.oecd.org/dataoecd/23/54/39558011.pdf>, retrieved October 2010: OECD.
- Plint, A., Moher, D., Morrison, A., Schulz, K., Altman, D., Hill, C. and Gaboury, I. 2006. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia*, 185(5), pp. 263-267.
- Rasmussen, N., Lee, K. and Bero, L. 2009. Association of trial registration with the results and conclusions of published trials of new oncology drugs. *Trials*, 10(1), pp. 116.
- Rodrik, D. 2008. *The new development economics: we shall experiment, but how shall we learn?* Boston: Working Paper, Kennedy School, Harvard University.
- Simes, R.J. 1986. Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology*, 4(10), pp. 1529-1541.
- Sorensen, G., Emmons, K., Hunt, M.K. and Johnston, D. 1998. Implications of the results of community intervention trials. *Public Health*, 19(1), pp. 379-398.
- Stewart, R., van Rooyen, C., Dickson, K., Majoro, M. and de Wet, T. 2010. *What is the impact of microfinance on poor people? A systematic review of evidence from sub-Saharan Africa*.

London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

What Works Clearinghouse 2010. Washington D.C.: What Works Clearinghouse. Available from: <http://ies.ed.gov/ncee/wwc/references/registries/index.asp?NoCookie=yes> [Accessed 29 October 2010].

WHO 2010. *WHO Registry Criteria (Version 2.1, April 2009)*. Geneva: WHO. Available from: http://www.who.int/ictrp/network/criteria_summary/en/index.html [Accessed 29 October 2010].

Zarin, D.A., Ide, N.C., Tse, T., Harlan, W.R., West, J.C. and Lindberg, D.A.B. 2007. Issues in the registration of clinical trials. *JAMA*, 297(19), pp. 2112-2120.

Zarin, D.A., Tse, T. and Ide, N.C. 2005. Trial registration at ClinicalTrials.gov between May and October 2005. *New England Journal of Medicine*, 353(26), pp. 2779-2787.